TELEPHONE THEORY AND PRACTICE

THEORY AND ELEMENTS

TELEPHONE THEORY AND PRACTICE

BY

KEMPSTER B. MILLER

THEORY AND ELEMENTS

MANUAL SWITCHING AND SUBSTATION EQUIPMENT

AUTOMATIC SWITCHING AND AUXILIARY EQUIPMENT

TELEPHONE THEORY AND PRACTICE

THEORY AND ELEMENTS

BY

KEMPSTER B. MILLER, M.E.
Fellow, American Institute of Electrical Engineers; Member, Western Society of Engineers; Author of "American Telephone Practice," 1904, and Joint Author of "Telephony," 1912

V. 1.

FIRST EDITION
FOURTH IMPRESSION

McGRAW-HILL BOOK COMPANY, INC.
NEW YORK AND LONDON
1930

PRINTED IN THE UNITED STATES OF AMERICA

PREFACE

It is now thirty years since the first edition of my "American Telephone Practice" appeared, and twenty-six years since its last revision. The day-to-day pressure of active professional work has prevented its being kept up to date in spite of frequent urgent appeals of my publishers and of an earnest desire on my part to do so.

When the present work was finally taken up in earnest, it was with the idea that it would be a complete revision of the old book under the old name "American Telephone Practice." It was soon realized, however, that this plan was impracticable. Since the last revision the telephone art had changed so radically and the industry had expanded so tremendously that to write about it, or even to think about it, in the old terms seemed inappropriate. Things like the automatic exchange system, which in 1904 were just beginning to appear as practicable to the bolder thinkers, had since become firmly established in the everyday workings of the industry. The even more revolutionary developments, like the three-electrode vacuum tube and the things that grew out of it, had not then even been dreamed of. More than all this, the very spirit of the times had changed. New conceptions of public service had crowded out the old. To have called a new work attempting to deal with this new order of things a revision of the old would have been akin to "the tail wagging the dog." Another consideration was the change that a quarter century of practical experience had wrought in the author himself. Whether for better or worse, even the old portions of the story could not be told in the same old way. And so, with real regret, it was decided to abandon the old name with the old work and let "American Telephone Practice" rest unchanged as a part of the records of telephony of the preceding generation.

In thus taking leave of this old friend I wish, at the same time, to express my deep appreciation of the reception which telephone people accorded it throughout its long life. The almost affectionate regard in which it seems to have been held has been the

main stimulus holding me to the determination to prepare, if possible, a successor in kind—one that would deal with present-day telephony in about the same simple way.

Like its predecessor, "Telephone Theory and Practice" is intended, not only for the student and the beginner in telephone work, but also as a general reference text for the more advanced. Perhaps also it may, in some degree, help the reader to coordinate in his mind the various phases of the industry that are so ably dealt with in the wealth of highly specialized literature made available by the developing, manufacturing and operating companies. As an industry grows in size and complexity and as its literature becomes more highly specialized, it becomes increasingly difficult for the man engaged in it to obtain a clear view of how his particular job is related to the rest of the business. A general work of this kind ought to assist in giving the needed perspective.

A single volume attempting to cover the subject of telephony in anything like a comprehensive way would be unwieldy. Moreover it would necessarily embrace more subject-matter than all would desire. The present plan, therefore, is to issue the work in several volumes, each covering, as far as possible, a distinct broad phase of the subject.

The first of these is the present volume, "Theory and Elements." It relates principally to underlying theory and attempts to lay a general groundwork for the more detailed discussion of telephone exchange systems, lines, systems of transmission and other subjects to be dealt with in succeeding volumes.

The science of acoustics, lying at the very root of the telephone transmission problem, has been given all too little attention in earlier general treatises. I have, therefore, included a rather full discussion of sound in both its physical and its physiological aspects in order to develop the fundamental requirements for the reproduction of speech and music. In developing the rudiments of the theory of multi-frequency alternating-currents, lines exactly parallel to those used in the discussion of sound waves have been followed with a view to bringing home to the student the fact that identically the same primary conceptions are involved whether dealing with waves of air or waves of electric current.

Like its predecessor, this work is mainly non-mathematical. Wherever mathematics has been used at all it goes no higher

than simple algebra or a mere suggestion of trigonometry. Even those without mathematical training should find most of it helpful, but those whose minds become blank at the mere sight of a mathematical formula are advised that they may skip it.

Acknowledgment has been made at appropriate points in the text of the sources of information on which I have drawn freely and also of the courtesy of the numerous companies who have furnished information and illustrative material. To all of these my thanks are due. I wish to express particular appreciation of the generosity of the Bell Telephone Laboratories in furnishing their many invaluable publications and in permitting the use of illustrations. Lastly my hearty thanks are due to my friend, Dr. Ralph D. Bennett of the Ryerson Physical Laboratory, University of Chicago, for his constructive criticism of much of my manuscript and his thoughtful reading of the proof. His knowledge of physical science and his clear thinking have helped me in the avoidance of errors and in the clearer statement of fact.

KEMPSTER B. MILLER.

PASADENA, CALIFORNIA,
July, 1930.

CONTENTS

PART I

INTRODUCTORY

The four chapters comprised within this part are intended to develop a sort of background for the beginner, to give him some conception of the broad significance of telephony, to make him somewhat familiar with the language of the art and to help him realize that the individual things treated of are not mere independent units but are intimately correlated with many other things which, all together, go to make up the telephone system.

TELEPHONE THEORY AND PRACTICE

THEORY AND ELEMENTS

CHAPTER I

MAN'S PROGRESS IN COMMUNICATION

We do not know, of course, when man began his efforts to communicate with his fellows. We are equally in the dark as to when language had its beginning. We do know, however, that the desire for mutual understanding is a predominant characteristic of man and has been the force impelling him to produce language.

We are told that the Neanderthal man, who lived in Europe during and after the fourth glacial period, was probably incapable of speech, as judged by the structure of his jaw bone. He, however, was one of the so-called "sub men," who were not the true ancestors of the human line. Some think that he was exterminated, perhaps thirty-five thousand years ago, by true man, who had no such structural speech limitation; but where true man came from and how far back *his* lineage goes we are totally uninformed. This is all hidden in the impenetrable darkness of countless thousands of years ago, and we can do no more than speculate about it.

Judging from the drawings made by true man on cave walls, probably thirty thousand years ago, and from the carvings and implements he left, we know that at that time he had risen, in artistic ability at least, far above the state of some savages to be found in some parts of the world to-day who have their own languages. We know, also, that among other things which distinguished him from the lower animals, he had a brain cavity of comparatively enormous size; and we have a right to assume

that he also had vocal organs capable of almost infinite development in tonal range and flexibility, and ear mechanism of such marvelous delicacy and perceptive power as to make it potentially capable of selectively sensing the entire range of vocal sounds. In view of all this it is difficult to believe that man had not felt the need for communication and developed some sort of a language long prior to this time.

At the outset primitive man probably communicated with his fellows, in so far as he communicated at all, by means of grunts or other elemental vocal sounds, grimaces, gestures and by direct touch. At first his sounds and signs probably did not mean much, but then he did not have much capacity either to think himself or to grasp the thoughts of others, so that these crude means of communication were probably quite sufficient. His interests were relatively few and mostly concerned with propagation and self-preservation; his emotions were elemental and vigorous, and his ideas infrequent and simple. He was undoubtedly gregarious, but social demands had not yet become exacting.

Language, wherein vocal sounds are definitely associated with ideas, probably did not exist at all. It seems likely that such vocal sounds as man made at first were merely for the purpose of attracting the attention of others, or for his own personal gratification, without any intent of communication. Song may have preceded speech by ages. Song need not have definite ideas attached; speech must have (or should have).[1]

With his splendid equipment of brain capacity, vocal organs and hearing mechanism, man subconsciously began to learn that articulate sounds afforded the best means of conveying and receiving definite ideas. The combination of voice and ear was incomparably more convenient and effective than any or all of the other means combined. The flexibility and modulating power of the voice permitted almost infinite variety of sound combinations to be produced. The ability of the ear selectively to sense these sounds was as great as that of the voice to produce them. Nothing approaching this effectiveness could be had through the sense of sight, touch, smell or taste. Moreover, the vocal means of communication did not interfere with the use of the hands, eyes or other members of the body while otherwise engaged and could be used under conditions where neither speaker

[1] JONES, R. L., The Nature of Language, American Institute of Electrical Engineers, *Transactions*, April, 1924.

nor hearer could see the other, an especially important consideration before the days of artificial fire and light.

Hence, as the ages progressed, sense became associated with sounds, words became significant, and, when combined into phrases and sentences, their significance was extended. Complex ideas could be expressed and understood.

So language has been subconsciously evolved, and, as we all know, the process is still going on. It has enabled man's power of expression to keep pace with his power of thought. It has advanced his ability to communicate from a point where he could convey a few elemental ideas by simple sounds and signs, or perhaps by a knock on the head with a club, to that where the now highly developed vocal and aural organs enable him to voice and to perceive almost every shade and subtlety of thought which the now highly developed brain is capable of harboring.

Just when man began to record his ideas we do not know. That is a question for the archaeologist or philologist to answer. We do know, however, that it was far back in prehistoric times, probably as much as twenty-five thousand years ago, as is shown by the records left by primitive men on the walls of caves and on flat pieces of stone. Whether the earliest attempts at drawing were intended to express definite meanings or were merely a sort of embryonic beginning of "art for art's sake" is a matter of doubt. It is certain, however, that, as in the case of sound, man began at an early date to associate sense or meaning with the various marks he could make on rock faces or on other surfaces. At first we think his writings took the form of mere ideagraphs, wherein a single symbol, probably suggestive in form, stood for a single name, object or idea, with no thought of connecting various symbols together, as in language. Later combinations of written symbols were used to convey more complex ideas. Also various written symbols became associated with definite sounds of spoken language, and, finally, written language was so highly developed that it could express nearly all of the thoughts and shadings permitted by spoken language.

With the development of parchment, papyrus and, finally, paper, man was enabled to lend an ease of handling and portability to his writings not possible with cave walls or stone slabs. Then the coming of the printing press enabled him to multiply indefinitely his singly handmade copies. His power to *reach out* with his ideas was thus enormously increased.

The struggle for increased distance of communication has been ever present. Long before historic times man could probably convey all the ideas he had, provided only his hearers were close enough to be within the sound of his voice. By beating on hollow logs or drums, he could transmit some ideas to greater distances within ear shot, and by gestures or other visual signals to greater distances within eyesight, but all of these efforts to communicate beyond direct voice range were at a great sacrifice in facility of expression.

Signaling by waving the arms was undoubtedly among the early efforts to communicate beyond the range of the voice. Later, after fire had been mastered, smoke signals by day and fire signals by night were used. All of these methods, depending upon direct visibility, have been used by savage and civilized men and to a limited extent are now used. The present semaphore of railway practice and the "wig-wag" flag signaling used in military operations are logical modern developments of man's early efforts at signaling by arm waving. The heliograph and the various systems of flash signaling are outgrowths of the simple signal fires of prehistoric times.

The employment of couriers, at first carrying word-of-mouth messages and, later, messages written on pieces of stone, bark, skin and, in modern times, paper, formed for ages practically the whole means of communication beyond the direct range of voice and eyesight. Our present world-wide system of mails is a development of this primitive method of communication, the supplanting of the courier in a large measure by railway, steamship, automobile and airplane, having, of course, enormously extended its effectiveness in range and speed. The mails as a means of communication leave little to be desired save in one often important respect—time.

Standing at the threshold of the nineteenth century and looking back as far as we can toward the obscure ages when man had this beginning, we find that marvelous progress has been made in communication in all respects save one: instantaneous distant transmission. In that respect he was about in the position of his Neanderthal precursor. For instant communication he was limited to distances within the sound of his voice, unless he resorted to louder sounds than those of speech or to visual signals, in either of which cases his range was somewhat increased,

but his power of expression was greatly curtailed by the necessity of using some prearranged code.

To illustrate: In 1775, under conditions demanding the greatest haste, what methods of communication were employed in touching off the American Revolution? Signal lights, code and word-of-mouth courier. According to song and legend, word that the British were coming, and coming by sea, was conveyed to Paul Revere across Boston Bay by lanterns hung "aloft in the belfry arch of the North Church tower as a signal light." The code was "One, if by land, and two, if by sea." Then, in turn, by his famous midnight ride, at the ponderous speed of horse-flesh, he conveyed the message "through every Middlesex village and farm" by pounding on doors and shouting. Truly archaic. Yet this was civilized man doing the best he knew how and using the most advanced means of that time. Fifty thousand years of progress had led to this!

It was not until the nineteenth century was well advanced that this standstill condition in regard to increased range of instant communication was broken and real progress began. The electromagnetic telegraph, coming about 1837, was the opening wedge. It was one of the outstanding achievements of all time and with its later developments, whereby several messages in both directions may be sent simultaneously over the same wire, or whereby many messages in all directions may be sent simultaneously without any wires, it has in itself revolutionized previous methods of fast, distant communication.

But even the telegraph requires the use of an elaborate code, which largely places it beyond the personal use of the communicators, and makes necessary the employment of skilled intermediaries—telegraphers or other operatives. Moreover, for its universal person-to-person application, the telegraph still clings to a vestige of the past in that the messenger boy—a survival of the romantic courier—is still largely used to gather the messages from their senders and deliver them at their destinations.

In 1875, just a century after the Paul Revere incident, the electric telephone was invented; and, on March 10 of the next year, it spoke its first sentence in an attic in Boston. A few days later the miracle of *distant* speech transmission was wrought. For the first time man could send his voice beyond the radius of a few hundred yards, which had been his limit for the countless ages of his existence.

This achievement is accredited to Alexander Graham Bell, a teacher of deaf mutes, a native of Scotland and a resident of the United States. Not the least astonishing feature of Bell's invention was the directness and the beautiful simplicity of his attack on the problem—so fundamentally sound that even instruments of most rudimentary character could produce the astounding result. Dom Pedro, then Emperor of Brazil, when it was exhibited to him at the Philadelphia Centennial in 1876, exclaimed, "My God, it talks." Sir William Thomson, later Lord Kelvin, the great British scientist, after seeing it at the same exposition, referred to "the hardihood of invention which devised such very slight means" to realize the desired end, and referred to the instruments themselves as of "quite a homespun and rudimentary character."

Bell gave the world scarcely more than an idea, but it was the correct idea, sound and vigorous. His instruments, feeble and crude to the point of calling forth skepticism and ridicule from the unknowing, served by their very feebleness and crudity to increase the profound admiration of the knowing. These instruments worked because they were vitalized by the spark of a great idea; an idea that was to revolutionize the ability of man to communicate with his fellows; an idea upon which the whole telephone industry, now grown to vast proportions, rests to-day, just as those two puny instruments in a Boston attic rested upon it fifty years ago. It was the beginning of a truly great art—telephony.

Bell's original telephone, now vastly improved, has been multiplied throughout the world about thirty-three million fold. His original fifty feet of line wire has grown until it now reaches all over the civilized world, having a length of over sixty-two million miles in the United States alone. Over fifty-nine million telephone conversations are now carried on each day in this country, and persons separated by the breadths of ocean and continent may now converse as clearly as if they were in adjoining rooms, their voices traveling with nearly the speed of light.

The Atlantic has been spanned and, as an early incident to this achievement, words spoken in the morning of one day on the Atlantic sea board were heard not only in Paris but also far out in the Pacific on the Hawaiian Islands in the evening of the calendar day before. Commercial telephone service is now continuously available between the entire United States

and the greater part of Europe. Ships far out at sea may converse with the mainland or with other ships; aeroplanes at dizzy heights may talk with each other or with stations on land; and, lastly, a speaker with ordinary voice may address at one time a hundred thousand persons gathered in one group about him, or millions scattered over the two hemispheres.

One must believe that this recently acquired power of man to transmit his thoughts in clear language and with instant speed throughout the length and breadth of land and sea will exert an even more profound influence on human relations as time goes on. It has been said that . . . "The capacity for free intercommunication between individuals of the species has meant so much in the evolution of man, and will certainly come in the future to mean so incalculably more, that it cannot be regarded as anything less than a master element in the shaping of his destiny."[1]

If ignorance and misunderstanding are among the great causes of human strife and unhappiness, then there appears no room for doubt that telephony, which has added so immeasurably to man's capacity for intercommunication, will exert a beneficent influence in disseminating knowledge and in helping people and peoples to understand one another. If it is true that the "screen of language" tends to create and maintain international hatreds, then this advance in the art of communication must tend to break down that screen. The telephone must ultimately exert a powerful influence on language itself, since, recognizing no boundaries, it brings peoples of different tongues into conversational contact.

But we must not expect results too soon. The telegraph is less than a century old, the telephone just over fifty and radio broadcasting about ten. The habits and prejudices that have been built up through the countless ages of man's existence cannot be wiped out or even greatly altered in any such short periods of time.

[1] TROTTER, I. W., "Instincts of the Herd in Peace and War"; see also presidential address of John J. Carty, delivered at Ninth Annual Meeting of the Telephone Pioneers of America.

CHAPTER II

PRECURSORS OF THE ELECTRIC SPEAKING TELEPHONE

MECHANICAL TRANSMISSION OF SOUND

The word "telephone" is derived from two Greek words, τῆλε, meaning *afar*, and φωνή, meaning either *voice* or *sound*. It is of interest to observe that the true derivation of the word does not necessarily carry into its meaning the idea of *voice* transmission. *Sound* transmission satisfies the etymological requirements nearly as well, and, in fact, it was with this meaning that the word "telephone" was first used. The word is considerably older than the electric speaking telephone, to which it is now almost exclusively applied.

When the speaking telephone did appear, it was at first commonly referred to as the "speaking telegraph," a misnomer, of course, but one naturally arising out of the fact that the electric telegraph had not, at that time, ceased to occupy the popular mind as a modern miracle. It is well to hold in mind these distinctions in terminology in briefly reviewing early developments in the art of sound transmission.

The need of increasing the range of voice transmission must be nearly as old as man's use of the voice for communication. Even today, when we shout to a person beyond our voice range, we feel this need, and primitive man, without any artificial aids, must have felt it more keenly. The Greeks tried to meet the need, when, before the walls of Troy, they employed Stentor "whose cry was as loud as the cry of fifty other men."[1]

But all such efforts involved only the straining of man's natural powers to their utmost. What was needed was artificial aid to accomplish the magic of distant speech transmission.

It seems natural that very early devices for providing artificial aid in speaking and hearing at a distance should have been in the nature of what we now call "speaking trumpets" and "ear trumpets"; instruments which serve more effectively

[1] "Iliad."

either to direct the voice toward the hearer or to concentrate the sound into the listener's ear. Devices of this nature, although they have been reinvented in modern times, probably date far back into antiquity. Indeed, it is difficult to believe that primitive man did not shout through his hands formed into a sort of trumpet, as we do to-day when we wish to send our voice to a distance beyond its natural range; or that he did not cup his hand to his ear, as we do to-day in an effort to increase the acuity of our hearing.

In his book on the history of inventions[1] Johann Beckmann 1739–1811), who has been referred to as the "founder of scientific technology," gives a chapter on Speaking Trumpets. This gives an early reference to "monstrous trumpets of the ancient Chinese, a kind of speaking-trumpets, or instruments by which words could not only be heard at the greatest distance possible, but could also be understood."

Beckmann, however, evidently did not place much reliance on these very ancient references, for, immediately after this reference to the ancient Chinese instruments, he states: "This invention belongs to the 17th century, though some think that traces of it are to be found among the ancient Grecians."

What we now know as an ear trumpet was evidently exhibited at the Royal Society in London in 1668 under the name "otacousticon." It was referred to by that most versatile of diarists, Samuel Pepys, in his entry of April 2, 1668, as follows:

"I did try the use of the Otacousticon, which was only a great glass bottle broke at the bottom, putting the neck to my ear, and there I did plainly hear the dancing of the oars of the boats in the Thames to Arundel Gallery window, which, without it, I could not in the least do."[2]

The speaking trumpet, as distinguished from the ear trumpet, came into prominence about 1670, two years later.[3] A controversy arose between rival claimants for the honor of its invention, in the course of which some one characterized it as a "new

[1] Beckmann, Johann, "A History of Inventions, Discoveries and Origins," 1780–1805; translated from the German by William Johnston, published by George Bell & Sons, London, 1884.

For this and many other references, see introductory chapter of "The Telephone and Telephone Exchanges," by J. E. Kingsbury, Longmans. Green and Co., London, 1915.

[2] "The Diary of Samuel Pepys."

[3] Kingsbury, *op. cit. p.* 2.

nicknamed old invention." This characterization was probably not inapt, particularly in so far as it related to the "nickname," for, in 1671, in a treatise on the invention, one of the claimants called it the "Tuba Stentoro-Phonica."

According to this account the largest one that he had employed at that time was 5 feet 6 inches long with a diameter of 21 inches at the large end and 2 inches at the small end. Of the trial he writes . . . "when by His Majesty's special command it was tried at Deal Castle by the Governor thereof, the voice was plainly heard off at Sea as far as the King's Ships usually ride, which is between two and three miles, at a time when the wind blew from the shore."

It is interesting to note, in connection witn this age-long development of the speaking- and ear-trumpet idea, that the "megaphone" as a device for aiding in hearing has, in our own times, been attributed to Edison.[1] What Edison did was to bring his genius to bear on the crudely worked out devices of ancient times. His own biographers[2] thus set forth his work in this line.

"The modern megaphone, now used universally in making announcements to large crowds, particularly at sporting events, is due also to this period as a perfection by Edison of many antecedent devices going back, perhaps, much further than the legendary funnels through which Alexander the Great is said to have sent commands to his outlying forces. The improved Edison megaphone for long-distance work comprised two horns of wood or metal about six feet long, tapering from a diameter of two feet six inches at the mouth to a small aperture provided with ear tubes. These converging horns or funnels, with a large speaking-trumpet in between them, are mounted on a tripod, and the megaphone is complete. Conversation can be carried on with this megaphone at a distance of over two miles, as with a ship or a balloon."

Regardless of what may have been the differences in form between the single straight-sided horn, which we now call the "megaphone," and the "Tuba Stentoro-Phonica" or other speaking trumpets of ancient times, it is probable that this three-horn combination of Edison represents the culmination

[1] See Megaphone, "Century Dictionary," New York.

[2] Dyer and Martin, "Edison: His Life and Inventions," Harper & Brothers, 1910.

of achievement in the speaking and ear-trumpet line up to that time.

As a modification of the speaking trumpet idea, one Captain John Taylor,[1] in 1845, invented an instrument "for conveying signals during foggy weather by sounds produced by means of compressed air forced through trumpets." No thought of transmitting speech was involved in this instrument, which merely produced powerful sounds derived from blasts of compressed air. Nevertheless, it was called "The Telephone" which is one of the early though by no means the first use of this word.

Another line of effort to extend the distance over which sounds could be sent, by direct transmission through air, was by means of what we now call the "speaking tube." This also is of ancient origin. Beckmann gives the following translation of a passage from della Porta, presumably from his "Magia naturalis," published in or prior to 1589.

"To communicate anything to one's friends by means of a tube. This can be done by a tube of earthenware, though one of lead is better . . . ; for whatever you speak at the one end the words issue perfect and entire as from the mouth of the speakers and are conveyed to the ears of the other, which in my opinion may be done for some miles . . . We tried it for a distance of two hundred paces, not having conveniences for a greater, and the words were heard as clearly and distinctly as if they had come from the mouth of the speaker."

In 1851 there was exhibited, at the London Exhibition, a speaking tube under the name of "telekouphonon." The same manufacturer also exhibited at that time what is thought to be a speaking trumpet, which he called the "Gutta Percha Telephone."[2]

All of the foregoing devices, whether speaking trumpets, ear trumpets or speaking tubes, worked on the principle of the direct transmission of sound through air. Another line of effort, distinctly a part of the quest for increased distance of sound transmission, was concerned with those devices which employed some solid medium, rather than air, as the vehicle for the transmission of the sound waves.

[1] Kingsbury, *op. cit.*, also "Year Book of Facts in Science and Art," p. 55, 1855.

[2] Kingsbury, *op. cit.* p. 4.

It has long been a matter of common knowledge that a light scratching or tapping on one end of a beam of wood may be distinctly heard by a person whose ear is pressed against the other end. Also, in modern times, that by placing one's ear against the rail of a railroad track, the sound of an approaching train may be heard at a greater distance through the rail than through air—and that the sound travels faster through the rail than through air. Beckmann says that knowledge of the transmission through a beam of wood "was known as far back as Pliny's time" (A. D. 23–79).

Dr. Robert Hooke, an English philosopher of astounding versatility, in the preface of the first edition of his "Micrographia," published in 1665,[1] makes the following quaint statement concerning the propagation of sound waves through bodies other than air and, particularly, through a wire:

"The next care to be taken, in respect of the Senses, is a supplying of their infirmaties with *Instruments*, and, as it were, the adding of *artificial Organs* to the natural; this in one of them has been of late years accomplisht with prodigious benefit to all sorts of useful knowledge, by the inventing of Optical Glasses.

* * * * * * *

"And as *Glasses* have highly promoted our *seeing*, so 'tis not improbable, but that there may be found many *Mechanical Inventions* to improve our *other* Senses, of *hearing*, *smelling*, *tasting*, *touching*. 'Tis not impossible to hear a *whisper* a *furlong's* distance, it having been already done; and perhaps the nature of the thing would not make it more impossible, though that furlong should be ten times multiply'd. And though some famous Authors have affirm'd it impossible to hear through the *thinnest plate* of *Muscovy-glass;* yet I know a way, by which 'tis easie enough to hear one speak through a *wall a yard thick*. It has not been yet thoroughly examin'd, how far *Otocousticons* may be improv'd, nor what other wayes there may be of *quickening* our hearing, or *conveying* sound through *other bodies* than the *Air:* for that that is not the only *medium* I can assure the Reader, that I have, by the help of a *distended wire*, propagated the sound to a very considerable distance in an *instant*, or with as seemingly quick a motion as that of light, at least incomparably swifter

[1] The entire preface to "Micrographia" is reprinted in the "Posthumous Works of Robert Hooke," London, 1705.

than that, which at the same time was propagated through the Air; and this not only in a straight line, or direct, but in one bended in many angles."

The spelling, punctuation and italics in this are given exactly as they appear in the original. This is probably the first suggestion of sound transmission through a wire. It is also perhaps the earliest recorded observation of the fact that sound travels through a solid faster than it does through air. But, so far as I am able to learn, Hooke did not describe the "Otocousticon" nor did he give any information about the kind of instruments, if any, he used at the two ends of his "distended wire." More will be said of this later.

About 1821 Charles Wheatstone, an English physicist, a man of great ingenuity and learning and, later, of large scientific achievement, took up the work of sound transmission through solids. Wheatstone began his career as a musical instrument maker and, through this practical work supplemented by ingenious and persistent experimentation, became well versed in acoustics. Later he turned his attention to electricity. The problem of the electrical transmission of intelligence became strong in his mind, and he was subsequently knighted for his work on the electric telegraph. He is perhaps best known among American telephone workers by association with the Wheatstone bridge, an instrument which, by the way, he did not invent,[1] although it almost universally bears his name.

As Kingsbury points out,[2] Wheatstone, of all the men who preceded Bell, was probably the man best fitted to have invented the electric speaking telephone. He had both acoustical and electrical knowledge, a rare combination. He also had an ingenious mind and a strong urge toward the transmission of intelligence. Moreover, he was well equipped for experimental and development work. Apparently, however, the electrical transmission of speech did not occur to him, and, while he did dream of its mechanical transmission, he could not rid himself of the idea that speech transmission (as distinguished from sound transmission) would necessarily involve vastly complicated mechanisms. As it was, his actual achievements in sound transmission went no further than that from room to room.

[1] Wheatstone's Bridge, "Encyclopedia Britannica"; also Wheatstone's Scientific Papers.

[2] KINGSBURY, *op. cit.*, p. 13.

In his earliest experiments Wheatstone observed that the sound vibrations of a tuning fork could be transmitted to a sounding board through a glass rod five feet long. Later, he transmitted "through rods of much greater lengths and of very considerable thicknesses, the sounds of all musical instruments dependent on the vibrations of solid bodies, and of many descriptions of wind instruments."[1]

This line of experimentation led him to the production, about 1821, of what he called "the enchanted lyre."[2] This consisted of a sounding box fantastically fashioned to resemble a musical instrument of classic shape. It was connected by a wooden rod with a piano located in another room, out of sight and hearing. In other cases the lyre was suspended by a wire from a piano located in a room above. However connected, the sound vibrations set up by the piano were transmitted to the lyre through the solid material of the rod or wire. This caused the lyre to give forth music as though originating within itself, much to the wonderment of the audience.

Wheatstone had some appreciation of the importance of speech transmission as well as of the difficulties of its accomplishment, as the following quotation from his paper[3] will show:

"The transmission to distant places, and the multiplication of musical performances, are objects of far less importance then the conveyance of the articulations of speech. I have found by experiment that all these articulations, as well as the musical inflexions of the voice, may be perfectly, though feebly transmitted to any of the previously described reciprocating instruments by connecting the conductor, either immediately with some part of the neck or head contiguous to the larynx, or with the sounding board to which the mouth of the speaker or singer is closely applied. The almost hopeless difficulty of communicating sounds produced in air with sufficient intensity to solid bodies might induce us to despair of further success; but could articulations similar to those enounced by the human organs of

[1] Wheatstone's Scientific Papers.

[2] New Experiments on Sound, by Wheatstone in Thomson's "Annals of Philosophy"; also, for suspended type of lyre, see article on Charles Wheatstone, "Encyclopedia Britannica."

[3] On the Transmission of Musical Sounds through Solid Linear Conductors and on Their Subsequent Reciprocation, *Journal of the Royal Institution*, 1831.

speech be produced immediately in solid bodies, their transmission might be effected with any required degree of intensity. Some recent investigations lead us to hope that we are not far from effecting these desiderata; and if all the articulations were once thus obtained, the construction of a machine for the arrangement of them into syllables, words, and sentences would demand no knowledge beyond that we already possess."

Apparently, therefore, he did succeed to the extent of feebly transmitting articulate speech mechanically; but his reference to the "almost hopless difficulty" of communicating the sound vibrations in air to solid bodies led him not to overlook but to abandon the simple course of speaking directly against a diaphragm. Instead, he proposed the astoundingly difficult course of producing sounds *similar* to those of human speech "immediately in solid bodies" in order that their transmission might be effected with sufficient intensity. Evidently, he proposed to simulate articulate speech by some sort of a machine that would, in itself, originate the various sounds of speech, and, having done this, to construct a machine that would break up and arrange these sounds into syllables, words and sentences. What he proposed was a machine that would originate speech, not one that would reproduce it as the phonograph does. Needless to say, neither Wheatstone nor anyone else has ever shown how such a machine, that could originate more than a few speech sounds, could be constructed.

It has been supposed that Wheatstone called his enchanted lyre "the telephone" as early as 1821, and that this was the earliest use of the word. Nothing however can be found in his writings or those of his associates that discloses the use of the word "telephone" before 1840. Then it was frequently applied to various devices for the mechanical transmission of sound. This use in 1840 was not, as Kingsbury surmises,[1] the earliest occurrence of the word "telephone," for as shown in Frederick L. Rhodes' recent work[2] it is of much older origin. It is to be found in a very small and rare book by G. Huth, Berlin, 1796, located in the library of Hamburg. This, in naming a plan of relaying spoken messages by speaking and ear trumpets, says: "What could be more appropriate here than the word derived

[1] Kingsbury, *op. cit.*, p. 10.

[2] Frederick L. Rhodes, "Beginnings of Telephony," p. 225, Harper & Brothers, 1929.

from the Greek: Telephone or Fernsprecher?" ("Telephon, oder Fernsprecher").

Wheatstone's transmission of sound through rods and wires leads naturally to a consideration of the device commonly known several decades ago by such names as "the lovers' telegraph," "the string telephone," the "mechanical pulsion telephone," etc. This in its simplest form consisted of two tin cans, the bottoms of which were replaced by a tightly stretched diaphragm of bladder or parchment, the centers of which were connected by a tightly drawn string. It is sufficiently illustrated

FIG. 1.—The lovers' telegraph.

in Fig. 1, which is taken from an early work on the electric telephone by Count Du Moncel.[1]

This simple device, had Wheatstone known of it, would have met at once "the almost hopeless difficulty of communicating sounds produced in air with sufficient intensity to solid bodies." He was on the verge of it when he connected his transmitting rod or "conductor" with "the sounding board to which the mouth of the speaker or singer is closely applied." But apparently his devices were too heavy, and he abandoned this line of attack.

As will be shown, Bell, when he attacked the problem of the electric telephone had a similar difficulty in that he could not,

[1] "The Telephone, the Microphone and the Phonograph," Harper & Brothers, 1879

at first, comprehend that such a simple thing as a thin membrane carrying a disc of iron could vibrate under the influence of the voice with sufficient intensity to produce the desired currents. Evidently Wheatstone did not know of the lovers' telegraph, and Bell in his testimony stated that he never heard of it until after the issuance of his patent.[1]

The origin of the lovers' telegraph is shrouded in mystery so far as the present writer is concerned, although others, at much earlier dates, have apparently felt no such uncertainty. In 1887, in his argument before the Supreme Court of the United States, Mr. J. J. Storrow, counsel for the Bell Company, in answer to a question of Mr. Justice Field as to when the string telephone was introduced, stated, "Two hundreds years ago it was described. It keeps disappearing and getting re-invented."[2]

Mr. Storrow undoubtedly had reference to the statement of Robert Hooke in the preface of his "Micrographia" already quoted in this chapter. Other authorities have stated without qualification that Dr. Hooke had reference to the string telephone.[3]

Neither in Hooke's "Micrographia," his "Posthumous Works" or his Biography has anything been found more definitely suggesting the lovers' telegraph than the matter herein quoted. Certainly this does not describe it, nor would it even suggest it to one not already familiar with it.

So far as Hooke's hearing of a whisper "a furlong's distance" is concerned, a speaking trumpet or an ear trumpet would have met his description equally well. He could scarcely have been referring to the lovers' telegraph as the means for making it "easie enough to hear one speak through a wall a yard thick," for even at this date it would puzzle one to use a lovers' telegraph for that purpose, without boring a hole through the wall, in which case a speaking tube would answer. As for the transmission through a wire, Dr. Hooke's description would be met by the scratching on one end of a wire, the other end of which was held to the listener's ear. Indeed, the fact that the Doctor

[1] "The Deposition of Alexander Graham Bell in the Suit brought by the United States to Annul the Bell Patents," p. 211.

[2] Oral Arguments for Bell Company in "The Telephone Appeals," p. 68.

[3] Preece and Stubbs, "A Manual of Telephony," p. 1, Whittaker & Co., London, 1893.

emphasized that his wire need not be straight, but might be "bended in many angles" is almost conclusive that he was not talking about anything like the lovers' telegraph, for, as we now know, unless they are slight and infrequent, bends are fatal to the successful performance of that instrument.

While Dr. Hooke did not describe the nature of the "Otocousticons" to which he alluded, some light is thrown on the subject by the reference in Pepys' diary, two years later, to the "Otacousticon," which was nothing more than an ear trumpet.

Hooke was probably the first to suggest the transmission of sound through a wire, and his is perhaps the earliest recorded observation of the fact that sound travels faster through a solid than through air—a real contribution to science. The diversity of his interests was amazing. They included such subjects as "Art of flying in the Air, and moving very swift on Land and Water," calculating machines, chronometers, navigation, astronomy and vivisection. He was truly a wonderful man, and hence it has been all too easy, with the knowledge of later years, to read into the vague passages of his writings more than he actually did, or meant. It is not inconsistent with his character to suppose that after making the suggestions, here quoted, in the preface of his book, he went no further in this line. It has been said in effect that his achievements might have been more profound had his interests been less diversified.[1]

Du Moncel, who wrote in 1879, at the time when the string telephone was claiming widest attention, stated:[2]

"It would be difficult to say with whom this idea originated, since it is claimed, as if beyond dispute, by several telephone-makers. If we may believe some travellers, it has long been used in Spain for the correspondence of lovers. However this may be, *it was not to be found among the scientific appliances of some years ago*, and it was even supposed by many persons that the cord consisted of an acoustic tube of slender diameter."

It is difficult to believe, if the string telephone had been of as great antiquity as many writers have considered it to be, it would not have been a common thing among the scientific appliances of the times immediately preceding the invention of the electric telephone. Great attention was being given to the

[1] Biography in "Posthumous Works of Robert Hooke."

[2] Du Moncel, "The Telephone, the Microphone and the Phonograph," Harper & Brothers, 1879.

science of sound, and this simple device would have been of great interest for its scientific value alone.

Certainly a familiarity with it would have enabled Wheatstone to have gone a step or two further than he did; and, as we shall see, it might have helped Bell over some difficult places in his work on the electric speaking telephone. But Bell did not know of it, and evidently Wheatstone did not—else why "the almost hopeless difficulty of communicating sounds produced in air with sufficient intensity to solid bodies?"

Contrary to the prevailing ideas as to the antiquity of the string telephone, I incline to the belief that it is comparatively modern, antedating the electric telephone (1875) by a few years at most. The earliest unmistakable description of it that I have been able to find is that of Adolph F. Weinhold,[1] Professor of the Royal Technical School, Chemnitz, Germany, in his book "Introduction to Experimental Physics," published in England in 1875. This reference was brought to my attention by the late Mr. Thomas D. Lockwood of Boston, whose remarkable memory was a veritable storehouse of telephonic lore.

Weinhold stated: "The transmission of sound can be strikingly shown by means of a tightly stretched piece of twine, or still better an iron wire. Each end of the cord or wire is fixed to the middle of a thin but not very small board called a *sounding-board,* which, in consequence of its comparatively large surface and great elasticity is peculiarly capable of receiving sonorous vibrations from the air, and, conversely, of communicating its own vibrations to the air." He described how the sound of a music box or of the voice could be thus transmitted over distances of more than 600 meters and with sufficient clearness to permit words to be understood and the characteristics of different voices to be distinguished. He also stated that "a short sharp cry" at one end could be heard twice at the other end, first by the more rapid transmission through the wire, and then by slower transmission through the air.

Regardless of its origin, the string telephone was a device of extremely limited application. Nevertheless, it was of scientific importance, since it taught at least two lessons in telephone fundamentals: first, that a thin diaphragm, such as a tightly stretched membrane, could take up the vibrations of the human

[1] "Introduction to Experimental Physics," p. 333, Longmans, Green and Co., London, 1875.

voice with sufficient intensity to pass them on to another solid body (the string); and, second, that so simple a thing as a diaphragm, when made to copy the original vibrations, would emit a substantial copy of the original sounds. We may profitably follow this device somewhat further to show that it was actually capable of limited useful application.

The most interesting example of an actually useful mechanical telephone that has come into my own personal experience was one connecting the houses of a father and his son, located about half a mile apart in the outskirts of Vineland, New Jersey. For a period of four or five years, from about 1885 on, this was in daily practical use for intercommunication between the two families.

The telephone instrument at each house consisted of a skin drumhead stretched tightly over a circular opening, about a foot in diameter, in a flat board about eighteen inches square. This board was rigidly and permanently mounted within the house on the outside wall nearest to the other house. Its top was tilted slightly inward from the vertical, and its center was perhaps seven feet above the floor line.

A small copper or bronze wire, about No. 16 B. & S. gage, was used instead of the string of the ordinary lovers' telegraph. This led into each house through an auger hole, bored through the outside wall, exactly opposite the center of the drumhead. It passed through the drumhead and terminated in an ordinary coat button of bone, about three-fourths of an inch in diameter. It was stretched fairly taut from the diaphragm in one house to that in the other and was supported throughout its length on poles set about one hundred feet apart. The wire was hung from each pole by a loop of cord about a foot long, and, as the poles were purposely set slightly out of line with each other and the wire hung from the inside of the bend, it always swung clear of the poles. The first point of support outside each house was carefully placed so that the wire would pass through the auger hole without touching. Thus, from one diaphragm center to the other, the wire swung clear of all rigid objects. The tension of the wire was, of course, taken by the button bearing on the center of each diaphragm, which was thus pulled an inch or more out of its normal plane.

This simple arrangement afforded, without any auxiliaries, not only means for talking but also for signaling; a problem

that was not so easily met in the development of the electric telephone. To use the instrument one stood on a chair and signaled to the other house by tapping with a knife handle, or similar object, on the button before him. The signal thus transmitted was sufficiently loud to be heard all over the premises of the other house. Upon response, the person at the called station also mounted a chair, and the two could then converse in ordinary tones, quite understandably. The transmitted speech, as may be surmised from our present knowledge, sounded like that of a person talking with his head in an empty barrel, but this did not seriously interfere with its usefulness.

This telephone, however, did have its drawbacks, not the least of which was the astounding noises that occurred when a bird flew against the wire, which happened frequently. The noises thus produced, while not as loud as thunder, were even more startling because they came with no warning. They would reverberate from one house to another in resounding thumps as the main wave caused by the impact of the bird would pass back and forth between the stations.

The mode of action often attributed to these mechanical telephones, that the wire or string moves bodily back and forth in the direction of its length in response to the pulls reciprocally exerted on it by the diaphragm, is, of course, in error, being contrary to the well-established theory of the transmission of sound waves through solids. What actually happens is that alternate waves of condensation and rarefaction follow each other through the material of the wire, just as they do through the material of air. The principal differences are that, in the case of the string or wire, the waves are confined to a restricted path instead of spreading out in all directions as in open air; and that, in the case of the copper wire, they travel about ten times as fast as they do in air.

The fact that the mechanical telephone has long since practically passed out of existence as a useful device (completely so far as I am aware), and that the electric telephone has increased by leaps and bounds, both in numbers and in usefulness, until to-day there are about 33,000,000 of them in daily operation, serves but to exemplify the vast fundamental difference between the two things. The mechanical telephone, depending on the actual transmission of the sound waves themselves, was naturally limited to short distances, and, by its very nature, its general

adoption in the field of public utility was an impossibility. On the other hand, the electric telephone, when it came, knew no such limitations of distance or congestion or general adaptability. It was a long step indeed from the highest development of the mechanical telephone to the electric telephone, as the next chapter will show.

CHAPTER III

EARLY HISTORY OF THE ELECTRIC SPEAKING TELEPHONE

The birthday of the telephone, in 1875, is only yesterday in human history; but it was early in the history of the application of electricity to the useful purposes of man. Electrical science itself had not long since begun to merge from the vague ideas that characterized the early part of the nineteenth century into the more definite and accurate knowledge of the twentieth century.

The telephone, of course, is an electromagnetic instrument, and the whole telephone art to-day is essentially an electromagnetic art. It is significant, therefore, to observe that it was not until 1819, a little over a hundred years ago, that Hans Christian Oersted, Professor of Natural Philosophy in the University of Copenhagen, discovered that there was a relationship between electricity and magnetism. Both electric and magnetic phenomena were known in a vague way before that time. Franklin had demonstrated that lightning was an electric manifestation; Volta had given the world the electric pile; and such magnetic actions as those of the lodestone and the mariners' compass had been observed from ancient times. But, although it was suspected, no one demonstrated that the two were kindred phenomena until Oersted observed[1] that a magnetic needle tends to place itself at right angles to a wire carrying a current of electricity. This was a simple discovery, as we look at it today, but it was one of epoch-making importance. It immediately stimulated research into electric and magnetic phenomena, paving the way for the revolutionary changes which electricity has since wrought in human affairs.

Ampere immediately took up the subject and in a very short time, with masterly thoroughness, formulated the laws upon which the present electromagnetic theory is based. Soon after Oersted's discovery, D. F. J. Arago, a Frenchman, and Sir

[1] "Annals of Philosophy," p. 273, 1820.

Humphrey Davy, celebrated English chemist, independently discovered the power of an electric current to magnetize iron and steel. It was William Sturgeon, however, who in 1824 made an electromagnet and called it by that name. His magnet was formed by wrapping a bare copper wire around an iron rod which had been bent into the form of a horseshoe and insulated with varnish. The convolutions of the wire were so spaced as not to touch each other.

A little later, Joseph Henry, who has recently been called the "Dean of American Scientists," made his classic experiments on the electromagnet.[1] To him must be accredited a large amount of our knowledge regarding it. He showed that it was better to insulate the wire itself rather than the core, and, in order to do this, he gave us silk-insulated magnet wire. He showed that best results could be secured by wrapping the wire closely along the whole length of the core and in successive layers, taking care to insulate the layers from each other by an intervening wrapping of silk ribbon. He also developed the method of winding the wire on the spool, and then slipping the spool on the core, a practice largely used to-day. In fact, Henry gave us the electromagnet of to-day except for improvements in the technique of its manufacture. Several of his magnets are now in the physics museum at Princeton University.

By these methods Henry constructed large magnets of great lifting power. Of more importance to us, however, were his so-called "intensity" magnets. These were small magnets wound with many turns of fine wire and capable of being operated over considerable lengths of line.

Early in 1831 he arranged a small office bell in such a manner that it could be tapped by the polarized armature of one of his "intensity" magnets. The coil of this magnet was connected in circuit with a mile of insulated copper wire suspended about one of the rooms of his academy. This was the first instance of magnetizing iron at a distance, or of a combination of magnet and battery so arranged as to be capable of such action. It was, therefore, the earliest example of an electro magnetic telegraph, all preceding experiments to this end having been on the galvanometer or needle principle.

[1] Article on Joseph Henry, "Encyclopedia Britannica." Bancroft, Gherardi, and Robert W. King, Joseph Henry, *Bell System Technical Journal*, January, 1926.

Henry thus blazed the trail leading to Morse's electromagnetic telegraph. He also furnished the electromagnetic essentials for Bell's telephone. In addition to this, as will be shown, he gave valuable furtherance to the invention of the telephone by showing kindly interest and advising a struggling young inventor who had come to him for assistance.

Oersted, Ampere, Arago, Davy, Sturgeon and Henry, in the work leading up to the electromagnet, established primarily one of the laws concerning the relationship between electricity and magnetism: that a current flowing in a conductor causes a field of magnetic force to exist about that conductor. This field of force may be considered as represented by "lines of force" encircling the conductor.

The electromagnet was the logical outcome of this law, since, by coiling the conductors many times about a core of iron, the fields of force due to each of the many convolutions were brought cumulatively into a comparatively small space, resulting in a greatly intensified field. The lines of magnetic force thread through the core and return through the air or other external path, always forming closed loops.

In 1831 Michael Faraday and Joseph Henry independently discovered the converse of these laws, relating to the transformation of electrical energy into magnetic. That is, they discovered that magnetic action could be transformed into electrical. They found that currents would be caused to flow in a closed conducting loop if the average intensity of the magnetic field passing through that loop was changed. The current so induced in the loop would flow only while such change was taking place, its strength would be proportioned to the rate of the change, and its direction would depend on the direction of the field and whether it was increasing or decreasing. It mattered not whether the change in the number of magnetic lines passing through the loop was caused by changes in the field itself or by a movement of the loop with respect to the field or by a movement of the field with respect to the loop. These laws of electromagnetic induction directly paved the way for the electric dynamo, just as the converse laws had paved the way for the electric motor.

These laws concerning the transformation of electric energy into magnetic and, conversely, the transformation of magnetic energy into electric are certainly the most important in the whole

realm of electrical science. Singly or together they form the foundation not only of the telephone and telegraph but also of electric lighting, electric power, radio and of every other achievement by which electricity has revolutionized the methods of life throughout the civilized world.

It was during the period of the discovery and formulation of electromagnetic laws that Dr. Georg Simon Ohm, in 1827, introduced a clear conception as to the relationship between the current, electromotive force and resistance in a circuit. This, when formulated, became Ohm's Law which has immortalized his name.

It was natural that first efforts toward the transmission of intelligence by electricity should have been in the direction of signal transmission as distinguished from speech transmission.

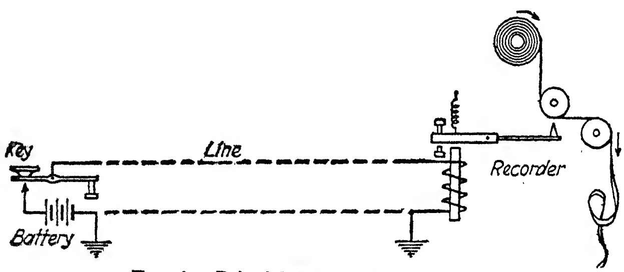


Fig. 2.—Principles of Morse telegraph.

To have aimed at the latter would have appeared then, as it did later, as a most visionary and foolish undertaking and, by some, as "wickedly defiant of God and His laws." Without overlooking the work of Wheatstone in England, already alluded to, and with due regard to that of Joseph Henry in preparing the way, it may be stated that the first practical electromagnetic telegraph was due to Samuel Finley Breese Morse, of Charlestown, Massachusetts, who was, at first, an artist of some note and then an inventor of lasting fame.[1]

Morse placed at one end of the line (Fig. 2) the electromagnet conceived by Sturgeon and developed by Henry and at the other end a battery and a key for opening and closing the circuit. Opposite one pole of his electromagnet he mounted a pivoted armature having a retractile spring with sufficient pull to hold

[1] Prime, S. I., "Life of Samuel F. B. Morse, LL.D.," D. Appleton & Company, New York, 1875.

the armature away from the magnet when no current was flowing. The free end of the armature carried a pen or stylus arranged to be brought into contact with a constantly moving strip of paper when the armature was attracted. By opening and closing the key the armature at the distant point could be caused to move up and down and thus record on the strip of paper the dots, dashes and spaces of the Morse code. The original Morse receiving instrument, now in the Historical Museum of the Bell Telephone Laboratories, is shown in Fig. 3.

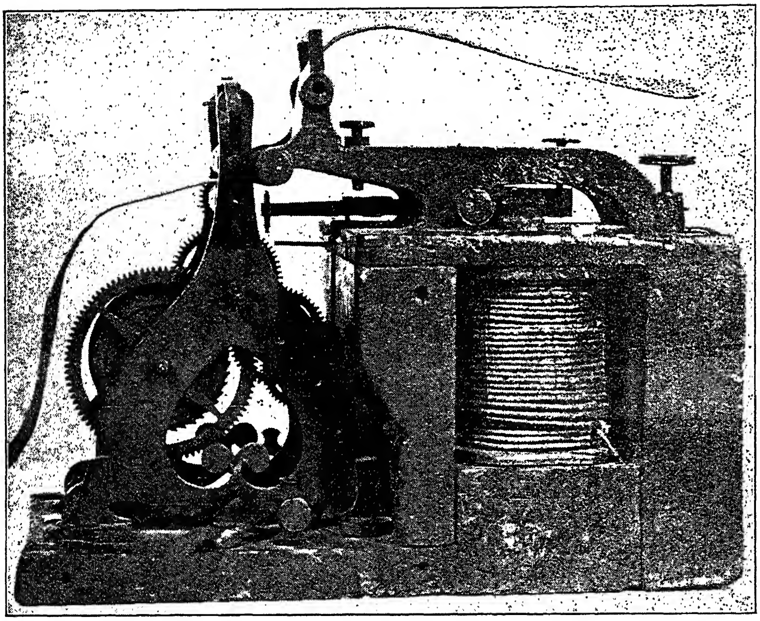


FIG. 3.—Original Morse telegraph recorder. (*Courtesy of Bell Telephone Laboratories.*)

It was not long after Morse's invention and its wide adoption that operators discovered that they could read the instruments by sound, a thing at first not believed possible. This led to the abandonment in most cases of the tape register. The receiving instrument was thus reduced to the very simple form of the well-known telegraph sounder shown in Fig. 4. This, in so far as its elements are concerned, consists of an electromagnet having a pivoted armature with a retractile spring and an anvil against which the armature lever will strike, both on its downward and upward motions.

The fact that in the Morse system of telegraphy the received signals are in the main read by sound does not mean, of course, that the system is in any proper sense one of sound transmission. While in a certain way the noise made by the sounder may correspond to the noise made by the key, this is purely incidental to the operation of the system. In fact, the noises of the sounder would occur just the same if the key were adjusted to operate without noise.

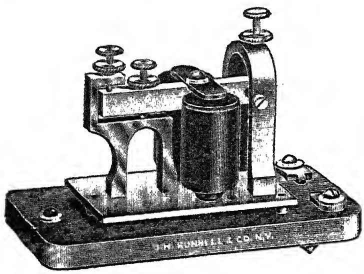


FIG. 4.—Modern Morse telegraph sounder

So far as extensive investigation by many searchers has shown, the first attack on the problem of transmitting speech to a distance by electricity was made by a Frenchman, Charles Bourseul, in 1854. Nearly all books relating to telephony have contained extracts from Bourseul's article,[1] but in view of its historical importance as being the first proposal of an electric speaking telephone and its general interest as containing early allusions to printing and autographic telegraphs, a translation is given here in full:

"The electric telegraph is based on the following principle: An electric current, passing through a metallic wire, circulates through a coil around a piece of soft iron which it converts into a magnet. The moment the current stops, the piece of iron ceases to be a magnet. This magnet, which takes the name of electromagnet, can thus in turn attract and then release a movable plate (plaque mobile) which by its to-and-fro movement produces the conventional signals employed in telegraphy. Sometimes this movement is directly utilized, and is made to produce dots or dashes on a strip of paper which is drawn along by clockwork. The conventional signals are thus formed by a combination of those dots and dashes. This is the American telegraph, which bears the name of Morse, its inventor. Sometimes this to-and-fro movement is converted into a movement of rotation. In that way we have either the dial telegraph used on railroads, or the telegraph used in the government system, which by means of two line wires and two indicating needles, reproduce all the signals of the aerial telegraph or semaphore

[1] *L'Illustration*, vol. 24, Paris, August 26, 1854.

which was formerly used. Suppose, now, that we arrange upon a movable horizontal circle letters, figures, signs of punctuation, etc. One can understand that the principle we have stated can be used to choose at a distance such and such a character, and to determine its movement, and consequently to print it on a sheet of paper appropriately placed for this purpose. This is the printing telegraph.

"We have gone still further. By the employment of the same principle, and by means of a mechanism rather complicated, it has been possible to reach a result which at first would seem to be almost a miracle. Handwriting itself is produced at a distance, and not only handwriting, but any line or any curve; so that, being in Paris, you can draw a profile by ordinary means there, and the same profile draws itself at the same time at Frankfort. Attempts of this sort have succeeded. The apparatus has been exhibited at the London Exhibition. Some details, however, remain to be perfected. It would seem impossible to go beyond this in the region of the marvellous. Let us try, nevertheless, to go a few steps further. I have asked myself, for example, if the spoken word itself could not be transmitted by electricity; in a word, if what was spoken in Vienna may not be heard in Paris? The thing is practicable in this way:

"We know that sounds are made by vibrations, and are made sensible to the ear by the same vibrations, which are reproduced by the intervening medium. But the intensity of the vibrations diminishes very rapidly with the distance; so that even with the aid of speaking tubes and trumpets, it is impossible to exceed somewhat narrow limits. Suppose that a man speaks near a movable disk, sufficiently flexible to lose none of the vibrations of the voice; that this disk alternately makes and breaks the connection with a battery; you may have at a distance another disk which will simultaneously execute the same vibrations.

"It is true that the intensity of the sounds produced will be variable at the point of departure, at which the disk vibrates by means of the voice, and constant at the point of arrival, where it vibrates by means of electricity; but it has been shown that this does not change the sounds. It is, moreover, evident that the sounds will be reproduced at the same pitch.

"The present state of acoustic science does not permit us to declare a priori if this will be precisely the case with syllables uttered by the human voice. The mode in which these syllables

are produced has not yet been sufficiently investigated. It is true that we know that some are uttered by the teeth, others by the lips, etc.; but that is all.

"However this may be, observe that the syllables can only reproduce upon the sense of hearing the vibrations of the intervening medium. Reproduce precisely these syllables.

"It is, at all events, impossible, in the present condition of science, to prove the impossibility of transmitting sound by electricity. Everything tends to show, on the contrary, that there is such a possibility. When the application of electromagnetism to the transmission of messages was first discussed, a man of great scientific attainment treated the idea as Utopian, and yet there is now direct communication between London and Vienna by means of a simple wire. Men declared it to be impossible, but it is done.

"It need not be said that numerous applications of the highest importance will immediately arise from the transmission of speech by electricity. Any one who is not deaf and dumb may use this mode of transmission, which would require no apparatus except an electric battery, two vibrating disks and a wire. In many cases, as, for example, in large establishments, orders might be transmitted in this way, although transmission in this way will not be used while it is necessary to transmit letter by letter, and to make use of telegraphs which require use and apprenticeship. However this may be, it is certain that in a more or less distant future, speech will be transmitted by electricity. I have made some experiments in this direction. They are delicate, and demand time and patience; but the approximations obtained promise a favorable result."

CHARLES BOURSEUL.

Paris, August 18, 1854.

To my mind his complete article makes Bourseul stand forth in a much more impressive light than do the mere extracts of it that are usually quoted. It shows that he had thought a long way into the problem of electrical speech transmission; that he saw, at least, some of the shortcomings of his proposed method; that he appreciated the transcendent importance of the achievement, when it should have been accomplished; and that he had complete faith in its ultimate consummation by some one. His idea of causing a distant plate to take up all of the vibrations of

another vibrating under the direct influence of the voice was a fine conception. But, as every experimenter during the next twenty years proved, the idea of accomplishing this by causing the original disc to *make and break* the circuit was all wrong. No one ever has succeeded in transmitting speech in that way. This was the stumbling block for the next two decades.

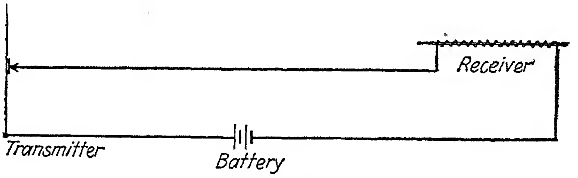


FIG. 5.—Principles of Reis' make and break telephone.

The next inventor to attack the problem was Philip Reis, a German physicist. In 1861 he constructed what he called a "telephone" and, during the succeeding period of several years, constructed several models, all exactly alike in principle. The theory of Reis' method may be understood from the diagram of

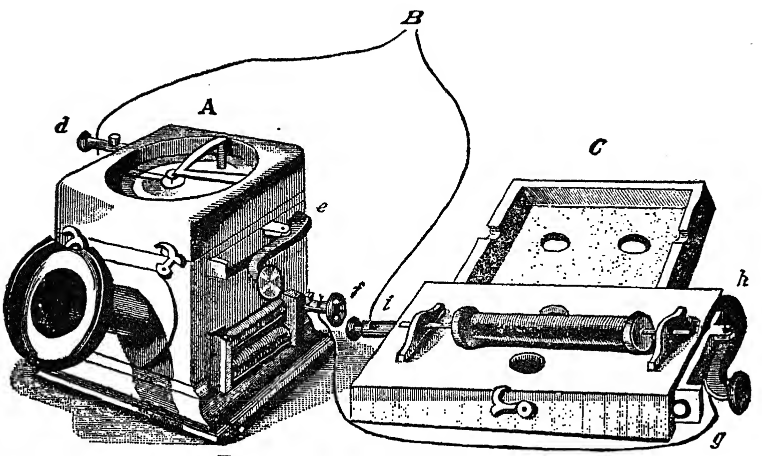


FIG. 6.—Set of Reis instruments.

Fig. 5, which shows the bare essentials of his instruments. Figure 6 shows one of his sets of instruments.[1] This consists of what he called the "telephone proper," at the left, and the "reproducing apparatus," at the right. In the "telephone

[1] U. S. *Reports*, vol. 126, p. 67, Supreme Court of the United States, October term, 1887.

proper," which was the transmitter, he employed a diaphragm consisting of a flexible membrane ("skin from the intestines of a hog") tightly stretched over a circular opening in a wooden box. To the center of this membrane was affixed a small plate of platinum connected by a flexible strip of metal to the binding post *d*. An angular piece of metal, shown above the diaphragm, carried a small platinum foot which rested on the platinum contact in the center of the diaphragm. This angular piece of metal, which formed a sort of rider on the diaphragm, was electrically connected to the telegraph key *e* mounted on the side of the box. From this the circuit continued through the telegraph sounder mounted just below the key and terminated on the binding post *f*. The receiver consisted essentially of a knitting needle around which was wound a coil of silk-insulated copper wire. This was mounted on a hollow sounding board or box so as to augment whatever sounds were given forth by the needle.

The action of the device, as described by Reis,[1] was as follows: "If now sufficiently strong tones are produced before the mouthpieces, their vibrations will put in motion the membrane and the angular little hammer (*winkelförmige Hämmerchen*) which lies on it; for every full vibration the circuit is once opened and again closed (*einmal geöfnet und wieder geschlossen*), and thereby are produced at station C in the core of the coil, just the same number of vibrations (*ebensoviele schwingungen hervorgebracht*) which are there perceived as tones or as combinations of tones (*accords*)."

It is seen that, so far as his transmitter was concerned, Reis followed exactly the path outlined by Bourseul—that of having the diaphragm, under the influence of the voice, make and break the circuit.

The action of his receiver was based on a phenomenon known as "Page's Effect." In 1837 Professor Page of Salem, Massachusetts, had discovered that a rod of iron suddenly magnetized or demagnetized would emit faint sounds. This was probably due to some sort of molecular or sub-molecular rearrangement causing slight changes in the dimensions of the bar, a phenomenon now known as "magneto striction." It was on these sounds, augmented by the sounding board, that Reis' receiver depended. The fact that telegraph apparatus was included by Reis in all his instruments is mute but eloquent testimony of the inability of

[1] *Ibid.*, p. 62.

these instruments to talk. This telegraph apparatus was, as stated in one of Reis' circulars, for convenience in experimenting.

Reis' telephone could be depended upon to transmit only the pitch of musical sounds. Obviously, it had difficulty in differentiating between sounds of different loudness, for a small break was likely to cause the same noise in the receiver as a large one—all either could do was to cause a complete cessation of the current. But what was even more important, the Reis instrument failed to transmit and reproduce the third and, so far as speech is concerned, the most vital characteristic of sound—its quality or timbre. As will be shown, quality or timbre depends neither on the pitch nor the loudness but on the number of overtones comprised in the sound and on their relative intensities with respect to the fundamental tone. Of course, the simple make and break of Reis' transmitter could not transmit quality because if the contact broke in response to the vibrations of the fundamental tone, it had no power to act in response to the more rapid vibrations corresponding to the overtones.

Reis' telephone attracted wide attention and was known among scientists of his time both in Europe and America. Numerous articles of that time describe it and its performance in detail. But not one of them even suggests any other mode of operation than that of making and breaking the circuit in accordance with the vibrations of the transmitter diaphragm. All agreed that the apparatus did reproduce melody—not, however, in the voice of the singer or of the musical instrument being used but in a voice all its own, which has been likened to the buzzing of an insect.

For something like a decade after Reis' efforts, no apparent advance was made in the art of electrical telephony. In 1874 Professor Alexander Graham Bell, then residing at Salem, Massachusetts, attacked the problem. Bell was a teacher of deaf mutes. He and his father and grandfather before him specialized in the science of sound and particularly in the art of teaching to speak those who were dumb because of having no sense of hearing. As a result of his training Bell had a greater knowledge of acoustics than of electrical science. It was probably this that led him to succeed where others had failed.

In 1874 Bell was working on a harmonic telegraph by which he hoped to send several simultaneous messages telegraphically over a single wire. He had strongly in his mind the problem of

speech transmission or, as he called it, the "transmission of vocal sounds by telegraph." The evidence shows conclusively that at this time Bell had conceived the correct method for the transmission of speech. He saw the reason for the failure of Bourseul and Reis and had it firmly in mind that, in the successful electric speaking telephone, the transmitter would be required to impress on the current, flowing in the line, undulations which would correspond in form to the motions of a particle of air engaged in the sound vibration.

Early in 1875 Bell visited Washington and called on Joseph Henry to seek his advice about his plan for "the transmitting of the human voice by telegraph." Henry showed much interest, said he though it was the "germ of a great invention," pointed out the difficulties in the way of its successful accomplishment, and when Bell told him that he thought he lacked the necessary electrical knowledge to overcome these difficulties, his laconic answer was "get it."

The encouragement thus given to Bell by the great scientist had a profound effect. The fact that such a man as Henry had not characterized his idea as visionary and impractical, as nearly everyone else had done, greatly increased his confidence.

The thought that a reed or disk forming the armature of an electromagnet might be made to vibrate in accordance with the sound of the voice and thus generate the sought-for undulating currents in the coil of the magnet had occurred to Bell. He felt confident that such a method, theoretically, would give him the desired currents undulating in conformity with the undulations of the sound waves. But he dismissed the idea because he thought the energy of the voice would be too feeble to generate currents sufficiently strong to result in sounds loud enough to be heard. Apparently, Henry in his interview had agreed with him that this method was correct in principle but probably impractical.

It was on June 2, 1875, that the feasibility of generating undulatory currents in this way was accidentally and dramatically demonstrated to Bell. On this day, he and his assistant, Thomas A. Watson, were experimenting on the harmonic telegraph in a hot garret in Boston which constituted his laboratory. In this harmonic telegraph a number of electromagnets of the kind shown in Fig. 7 were connected in series in the same circuit. Before the pole of each magnet there was a steel reed forming

the armature. Each receiver was tuned to vibrate at a frequency different from that of the others. The idea was that each would vibrate selectively in response to current of its own frequency impressed in the line by the transmitters. What happened on the occasion can best be told in Mr. Watson's own language.[1]

"I had charge of the transmitters, as usual, setting them squealing one after the other, while Bell was retuning the receiver springs one by one, pressing them against his ear as I have described. One of the transmitter springs I was attending to stopped vibrating and I plucked it to start it again. It didn't start and I kept on plucking it, when suddenly I heard a shout from Bell in the next room, and then out he came with a rush,

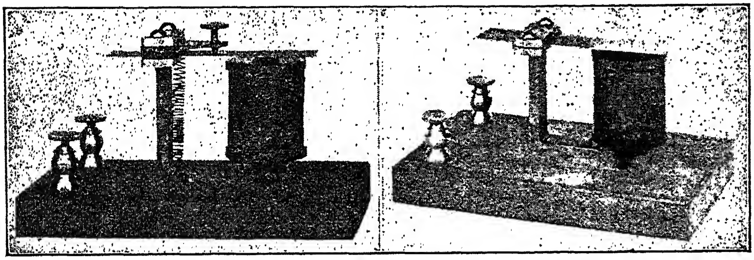


FIG. 7.—Bell's vibrating reed (transmitter left, receiver right). (*Courtesy of Bell Telephone Laboratories.*)

demanding, 'What did you do then? Don't change anything. Let me see!' I showed him. It was very simple. The make-and-break points of the transmitter spring I was trying to start had become welded together, so that when I snapped the spring the circuit had remained unbroken while that strip of magnetized steel by its vibration over the pole of its magnet was generating that marvelous conception of Bell's—a current of electricity that varied in intensity precisely as the air was varying in density within hearing distance of that spring. That undulatory current had passed through the connecting wire to the distant receiver which, fortunately, was a mechanism that could transform that current back into an extremely faint echo of the sound of the vibrating spring that had generated it, but what was still

[1] Birth and Babyhood of the Telephone, by Thomas A. Watson, address delivered before the Third Annual Convention of the Telephone Pioneers of America, Chicago, October 17, 1913.

WATSON, THOMAS A., "Exploring Life," D. Appleton and Company, New York and London, 1926.

more fortunate, the right man had that mechanism at his ear during the fleeting moment, and instantly recognized the transcendent importance of that faint sound thus electrically transmitted. The shout I heard and his excited rush into my room were the result of that recognition. The speaking telephone was born at that moment. Bell knew perfectly well that the mechanism that could transmit all the complex vibrations of one sound could do the same for any sound, even that of

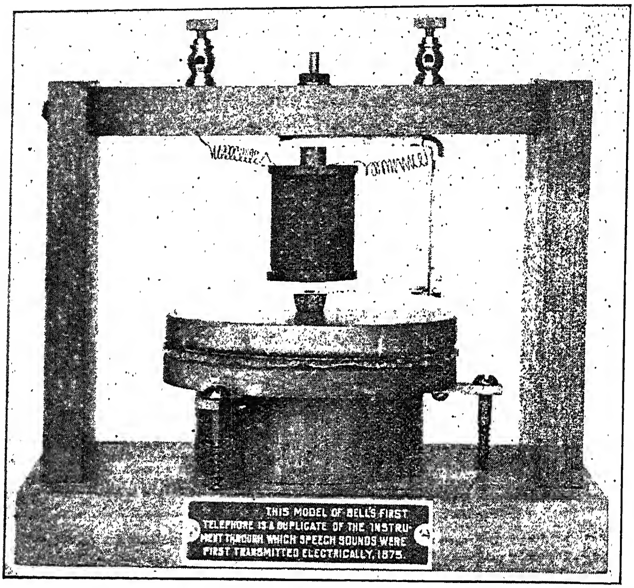


FIG. 8.—First electric speaking telephone. (*Courtesy of Bell Telephone Laboratories.*)

speech. That experiment showed him that the complex apparatus he had thought would be needed to accomplish that long dreamed result was not at all necessary, for here was an extremely simple mechanism operating in a perfectly obvious way, that could do it perfectly."

After many repetitions of this experiment with different reeds that same afternoon, Bell sketched out for Watson the plan for the telephone instrument he was to make. So simple was it in construction that Watson had it ready to test the next afternoon.

This, the first speaking telephone, is shown in Fig. 8. It consisted merely of an ordinary frame on which was mounted one of Bell's harmonic receivers like that of Fig. 7. A tightly

stretched parchment drumhead was mounted on this frame, the center of which was fastened to the free end of the receiver armature. The idea of this instrument was that the diaphragm or drumhead would take up the vibrations of the voice and force the armature into exactly similar vibrations. These, in turn, by electro-magnetic induction, would cause currents in the winding of the magnet that would undulate in conformity with the vibrations of the voice.

The receiver used with this first transmitter, according to Watson, was merely one of Bell's vibrating reeds (Fig. 7) held so closely against the ear of the listener as to damp its natural rate of vibration.

On the first trial,[1] the night of June 3, 1875, Bell could hear nothing, but Watson, on account of his more acute hearing and Bell's more vigorous voice, could unmistakably hear the tones of Bell's voice and "almost catch a word now and then." The trial was disappointing to Bell, but, nevertheless, it showed him that he was on the right track.

The building, 109 Court Street, Boston, in which these events occurred, is still standing (1930). On it has been placed a bronze tablet reading as follows:

HERE THE TELEPHONE WAS BORN, JUNE 2ND, 1875 THE BOSTONIAN SOCIETY AND THE NEW ENGLAND TELEPHONE AND TELEGRAPH COMPANY PLACED THIS TABLET MARCH 10TH, 1916

Being convinced that his principle was sound, Bell next devoted himself to the preparation of his patent application. This was filed in the United States Patent Office on February 14, 1876. So novel was the invention that the patent[2] issued only three weeks later. It was probably the most valuable patent ever issued in any art or in any country. It was entitled "Improvement in Telegraphy" and was directed not only to the "telegraphic" transmission of speech but of signals as well.

[1] Birth and Babyhood of the Telephone, by Thomas A. Watson, address delivered before the Third Annual Convention of the Telephone Pioneers of America, Chicago, October 17, 1913.

[2] U. S. Patent 174,465, March 7, 1876.

Figures 5 and 7 of the patent are reproduced as Figs. 9 and 10, respectively. Figure 9 (Fig. 5 of the patent) was used to show the general principle involved, whether used in the transmission of signals or of speech. It will be recognized as exhibiting exactly the conditions which existed in the reed plucking incident on June 2, 1875, at which time, as Watson says, "the birth cry

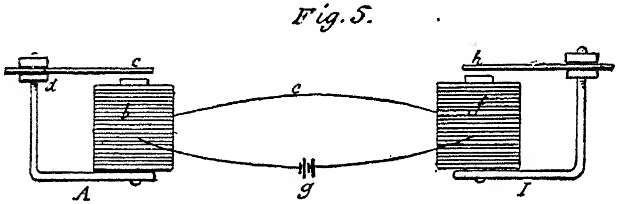


Fig. 9.—Figure 5 of Bell's patent.

of the telephone was heard." Two of the reeds, such as those of Fig. 7, are connected in simple series relation in circuit with a battery. If the reed *c* is made to vibrate from any cause, it will, by electromagnetic induction, cause undulatory currents to flow in the coil *b*. These will cause the magnet *f* of the distant instrument to exert a varying attraction and thus cause its reed *h* to vibrate.

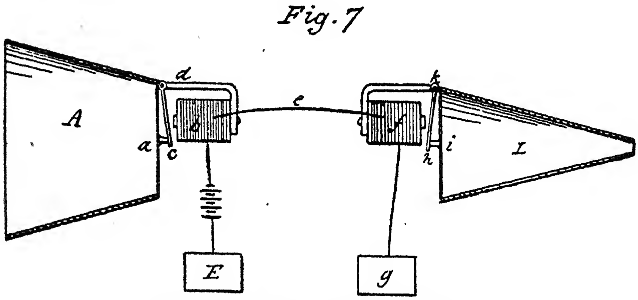


Fig. 10.—Figure 7 of Bell's patent.

This mode of operation, as applied particularly to voice transmission, is illustrated in Fig. 10 (Fig. 7 of the patent). At the transmitter, a diaphragm is tightly stretched over the small end of a speaking trumpet or mouthpiece, while, at the receiver, another diaphragm is similarly stretched over the large end of an ear trumpet or earpiece. To the centers of these diaphragms are attached the free ends of their respective arma-

tures. The sound vibrations of the voice are thus taken up by the transmitter diaphragm and imparted to its armature, while the similar vibrations produced in the receiver armature are imparted to its diaphragm, thus reproducing the sound. Bell realized that in speech transmission he should, as far as possible, get away from any fixed rate of vibration of his armatures, and, accordingly, in Fig. 7 of the patent, the armature of each instrument was loosely pivoted instead of being spring mounted.

The part of the patent specification relating particularly to the telephone occupied not more than about twenty-five lines of print, but it revealed Bell's grasp of the problem in hand. It disclosed not only the fundamental method, without which no one then or since has been able to transmit speech electrically, but it also gave a perfectly feasible plan of practicing that method. By it he told the world for the first time how to transmit speech electrically. Following its instructions any skilled electro-mechanic could have made and used the invention.

Even if, up to the time of filing his patent application, Bell's device had never spoken a word or if it never had had any experimental work done on it, this description alone would have sufficed. It is to the man who shows the world how to make and use an invention that the patent laws of the United States confer the honor and the reward.

While the accidental discovery at the time of the "reed plucking" incident showed Bell the feasibility of the magneto principle of operation, he, even then, realized that the action of the magneto transmitter would be feeble. His patent application clearly showed that he had other and more powerful methods in mind. Among these was that of having the transmitter vary the external resistance of a circuit containing a battery and receiver. Under this plan a battery, which could be made as strong as desired, would furnish the current, and the slight energy of the voice would operate a sort of valve to cause undulations in this current. Sometime before March 10, 1876, he worked on the practical development of this method, and one of his early embodiments of it in working form is crudely illustrated in Fig. 11, a sketch made by Mr. Watson and used in connection with his deposition in one of the numerous patent suits in which Bell's invention became involved.[1] Incidentally,

[1] General Brief for Bell Company, p. 472, Supreme Court of the United States, October term 1886.

this sketch is quite typical of the kind of sketches telephone men draw for each other today.

In accordance with Watson's description given at the time, the diaphragm against which the voice is to be directed is shown at *A*. To its center is attached a cork *B* from which depends the "plunging wire" *C* connected to the binding post *H*. The vessel *I*, of glass, is partly filled with acidulated water to the level *J*, and into this water the point of the plunging wire *C* dips "just enough to make sure that it will never vibrate out." Near the edge of the vessel is a stationary metal rod *G* immersed deeply into the liquid. This forms the other terminal of the instrument.

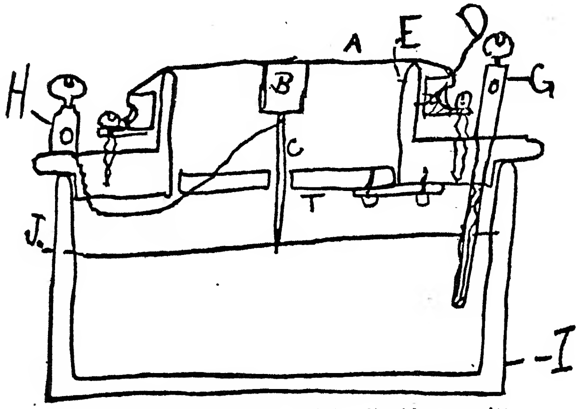


FIG. 11.—Watson's sketch of Bell's liquid transmitter.

The mode of operation was thus described: "The current, in passing from the rod *C to* the liquid, the upper surface of which is shown by the line *J*, experiences considerable resistance *at the contact of the rod and the liquid*, and that resistance is, other things being equal, proportionate to the area of contact. If that rod just dips into the liquid an extremely short distance, then its vibrational movement up and down will *notably* vary the area of contact, and consequently notably vary the electrical resistance at that point, and therefore the current which is affected by that resistance."

It was with an instrument of this kind that the next milestone in the history of the telephone was passed. With it the telephone spoke its first sentence. Evidently the achievement came rather unexpectedly, for, as Mr. Watson says, had he and Bell known that the instrument they were about to test was to prove the

best of any that had gone before, they would probably have rehearsed the event and decided upon some such impressive message as Morse's "What hath God wrought?" Instead of that the message was quite commonplace—merely the simple request, "Mr. Watson, come here. I want you." It is needless to say that Watson came.

Nearly forty years later, on January 25, 1915, on the occasion of the opening of the transcontinental line between New York and San Francisco, Bell, in New York, speaking, through an exact replica of his first instrument, said again to Watson, "Mr. Watson, come here. I want you." But this time Watson did not come so quickly, for he was over three thousand miles away in San Francisco.

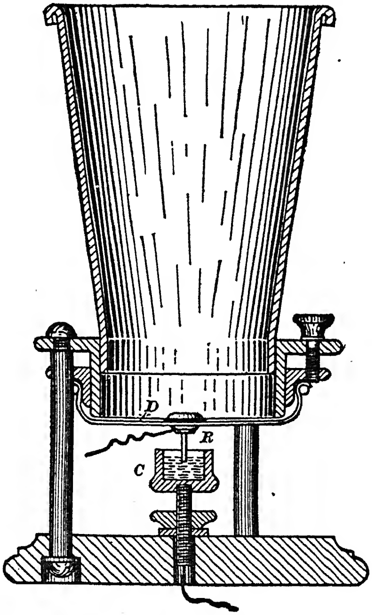


Fig. 12.—Bell's Centennial liquid transmitter.

This incident is parenthetically introduced at this point for two reasons: It shows how fundamentally perfect had been Bell's original conception; and it affords a striking illustration of the improvements that have been made in all phases of telephone transmission aside from those involved in the telephone instrument itself. The layman is likely to think of the telephone art as being represented by the instrument on his desk or wall, into which he talks and to which he listens. He does not realize that between him and the person with whom he is conversing there is a network of lines, switchboards and other instrumentalities which, taken together, form one of the most intricate and highly developed organizations yet conceived by man. This incident at the opening of the transcontinental line showed that, throwing away all of the advancement that had been made in the telephone instrument itself, the improvements in the line and other parts of the system were such as to enable an instrument, which, in Bell's time, had succeeded imperfectly in making itself heard over a line perhaps fifty feet

in length, to transmit understandable articulation to another instrument across the continent.

Bell decided to exhibit the telephone at the Centennial Exposition held in Philadelphia in 1876, and for that purpose he had Mr. Watson make additional models of all of the more successful instruments that he had used up to that time. These are shown in Figs. 12, 13 and 14 and require little further description.

In the Centennial model of the liquid transmitter (Fig. 12), the "plunging rod" of the earlier instrument was made of a short rod of carbon *R*, and the liquid was either of acidulated water or mercury held in a metallic cup *C*. The Centennial model of the magneto transmitter (Fig. 13) consisted of an electromagnet *H* in front of the core *C* of which was adjustably

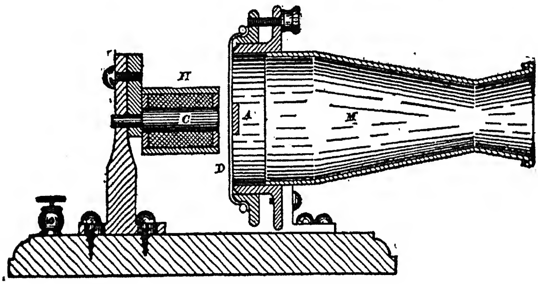


FIG. 13.—Bell's Centennial magneto transmitter.

mounted a diaphragm of goldbeater's skin *D* carrying a small iron armature *A* at its center. The Centennial magneto receiver (Fig. 14) consisted of a tubular magnet, composed of a coil *H*, surrounding a soft iron core *C* inclosed in an iron tube *E* which was about 1¾ inches in diameter and 3 inches long. This tube was closed by a thin iron armature or diaphragm *D* which rested loosely on the upper face of the iron tube. It was held in place merely by the magnetism of the core and was not secured at one edge by a small screw, in which form it is often but erroneously illustrated.

It was the performance of these instruments at the Centennial that caused the profound sensation that has been referred to in Chap. I. It was fortunate that on the board of awards of the exposition were two of the outstanding physicists of that time—Sir William Thompson, later Lord Kelvin, and Joseph Henry. The layman was inclined to scoff or to disbelieve, but these

men were inspired not only by the astounding performance but also by the simplicity of the means by which it was accomplished.

With all of his Centennial magneto instruments Bell employed a battery in the circuit to magnetize the cores and thus secure the magnetic fields in which the iron armatures vibrated. He had discovered, however, at the time of the reed plucking incident, that the instruments would work, through less powerfully, without any battery in the circuit.[1] This was true

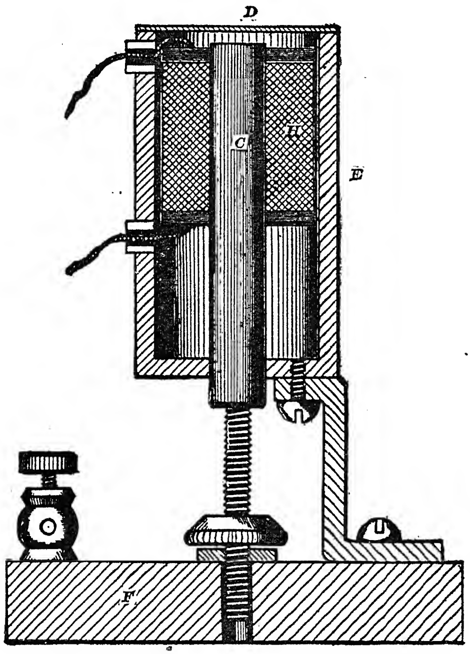


Fig. 14.—Bell's Centennial receiver.

because his iron cores, probably of none too good quality, had previously been strongly magnetized and retained enough residual magnetization to afford the necessary field. This, of course, led to the use of a strong permanent magnet instead of one that had been weakly magnetized merely as an incident of its earlier history. Then followed a period of experimentation during which all kinds and shapes of magnets and diaphragms

[1] General Brief for Bell Company, Supreme Court of the United States, October term, 1886, p. 52, letter from Bell to Gardiner G. Hubbard, June 2, 1875.

were tried and from which emerged the prototype of the modern telephone receiver. It was then used, however, as both receiver and transmitter. Two such instruments are shown diagrammatically in Fig. 15. They are connected together in the line circuit without any battery—the simplest form of telephone system. Each instrument is alternately transmitter and receiver. When transmitter, it is a dynamo, converting the energy of the voice into alternating currents; when receiver, it is a motor, converting these currents into useful work.

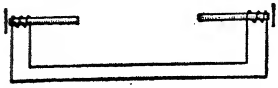


FIG. 15.—Two magneto telephones connected by line.

Bell's contribution to the progress of the world was succinctly set forth in his famous "fifth claim" which he himself drew, and which reads as follows:

"5. The method of, and apparatus for, transmitting vocal or other sounds telegraphically, as herein described, by causing electrical undulations, similar in form to the vibrations of the air accompanying the said vocal or other sound, substantially as set forth."

Probably no patent claim has ever been subject to so many and to such vigorous attacks as has this one. As soon as the importance of the telephone was realized by the public, numerous inventors claiming to have antedated Bell, and others, interested from other motives, launched suits for the purpose of breaking down this patent and, particularly, its fifth claim. There were hundreds of these suits, some of them finally being brought to the Supreme Court of the United States. Against all these attacks the patent was invulnerable. Every decision in every court from the lowest to the highest was in its favor, and, when the litigation had ended, Bell's position as the original inventor of the telephone had been established beyond dispute, and his patent had been adjudicated as broad enough to cover all known methods of transmitting speech by electricity.

The patent has, of course, expired, but it is of interest to note that no one, even to this time, has been able to suggest a method of transmitting speech electrically that did not fall within the scope of Bell's fifth claim.

No history of the early development of the telephone would be complete without some further allusion to these patent controversies. Of course, the early proposal of Bourseul and the later attempts of Reis to carry out Bourseul's teaching were

among the principal pieces of ammunition used by the attacking parties. It is perfectly clear in the light of present knowledge why the methods of Bourseul and Reis could not succeed. The "make-and-break" principle was obviously all wrong. Furthermore, Reis' receiver, depending on the "Page Effect," would have been most inefficient even if it had been coupled with a transmitter based on the correct principle.

Claims have been made by those who came after Bell that the Reis instrument could be made to talk. But in so far as this has ever been done, it was done by so adjusting the transmitter as to make it impossible for it to break the circuit at all. So adjusted it might operate very inefficiently as a variable resistance transmitter. This was not, of course, the teaching of Reis but of Bell. In view of our present-day knowledge, and after a review of the documentary evidence contemporary and prior to the time of this controversy, one can not fail to be impressed with the soundness of the action of the Supreme Court in disposing of the alleged anticipation by Bourseul, Reis, Van der Weyde and all those who followed the make-and-break principle.

The following is a part of the decision of the court[1] which finally disposed of the Reis matter:

"We have not had our attention called to a single item of evidence which tends in any way to show that Reis or any one who wrote about him had it in his mind that anything else than the intermittent current caused by the opening and closing of the circuit could be used to do what was wanted. No one seems to have thought that there could be another way. All recognized the fact that the 'minor differences in the original vibrations' had not been satisfactorily reproduced, but they attributed it to the imperfect mechanism of the apparatus used, rather than to any fault in the principle on which the operation was made to depend.

"It was left for Bell to discover that the failure was due not to workmanship but to the principle which was adopted as the basis of what had to be done. He found that what he called the intermittent current—one caused by alternately opening and closing the circuit—could not be made under any circumstances to reproduce the delicate forms of the air vibrations caused by the human voice in articulate speech, but that the true way was to operate on an unbroken current by increasing and diminishing

[1] U. S. *Reports*, vol. 126, p. 544, Supreme Court.

its intensity. This he called a vibratory or undulatory current, not because the current was supposed to actually take that form, but because it expressed with sufficient accuracy his idea of a current which was subjected to gradual changes of intensity exactly analogous to the changes of density in the air occasioned by its vibrations. Such was his discovery, and it was new. Reis never thought of it, and he failed to transmit speech telegraphically. Bell did, and he succeeded. Under such circumstances it is impossible to hold that what Reis did was

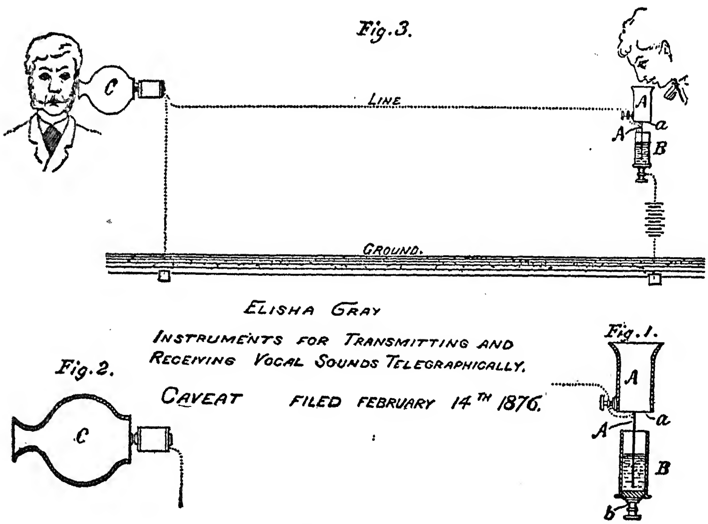


FIG. 16.—Drawing of Gray's caveat.

an anticipation of the discovery of Bell. To follow Reis is to fail, but to follow Bell is to succeed."

Another important contestant in this series of litigations was Elisha Gray. By a strange stroke of chance, Gray filed a caveat in the United States Patent Office a few hours later in the same day that Bell filed his application. The drawing in Gray's caveat is reproduced in Fig. 16. The intended mode of operation is obvious. Bell's liquid transmitter (Figs. 11 and 12) depended on variation in the extent of the immersion of the electrode, while Gray's instrument, owing to the great extent to which the vibrating electrode was immersed, depended rather on the variation in the length in the conducting path through the liquid itself, obviously a faulty principle for this purpose.

But, ignoring the question as to the practical operation of Gray's instrument, a caveat, by its very nature, is an admission by the inventor that his invention is not complete—that it is an idea upon which he intends to work further, and upon which he intends to file an application when it is sufficiently developed in his mind. Gray, like Bell, had been working on a harmonic telegraph and did produce a successful one. He was not diligent, however, in following up his idea for a telephone, for he dropped the matter for a number of years after filing his caveat. It was not until after the importance of Bell's telephone had been appreciated by the public that the Western Union Telegraph Company acquired Gray's alleged rights and opened fire on the Bell patents, claiming that Gray was the prior inventor. Somewhat later it transpired that the Western Union Company's own counsel, after a most extended study, advised his client that the rights of Gray could not possibly prevail as against those of Bell, and, on the basis of that advise, the Western Union Company settled the matter out of court.

Another principal contestant was one Daniel Drawbaugh of Eberly's Mills, a small village in Pennsylvania. Drawbaugh, a man of limited education but of considerable ingenuity and skill as a mechanician, was well known in his village circles. He claimed to have made successful electric speaking telephones several years prior to Bell's. A remarkable phase of this contest was that his interests brought forth many witnesses who testified that he had done so, and that they had actually heard his instruments speak. Although Drawbaugh had visited the Centennial and had seen Bell's instruments there, and although he had been aware of the fame that had come to Bell as a result of his invention, he apparently remained indifferent. It was not until several years had elapsed and after he had fallen into the hands of the promotors of a telephone company, which was started for the purpose of exploiting alleged inventions of its own, that Drawbaugh put forth his claim and appeared as a contestant.

The Supreme Court was divided on the question of Drawbaugh's rights. The majority decided against him however, and in favor of Bell on the general ground that it was not within the scope of conceivable human conduct for a man of Drawbaugh's intelligence, if he had made all of the telephones that he claimed to have made prior to Bell's time and had them in his shop at the time he visited the Centennial, to have remained silent for a

period of about four years thereafter. The court stated that they had not overlooked the testimony of the large number of witnesses to the effect that Drawbaugh made successful instruments but that the effect of this testimony had been completely overcome by the conduct of Drawbaugh from the time of his visit to the Centennial until he was put forward by the promotors nearly four years after. The court said further:

"The news of Bell's invention spread rapidly and at once, and it took but a few months to demonstrate to the world that he had achieved a brilliant success. If it were known at Eberly's Mills alone that Drawbaugh had been doing the same thing for years in his shop there—and it certainly would have been known all through the little village if it had actually been done—no one can believe that the public would be kept in ignorance of it until four years afterward . . ."

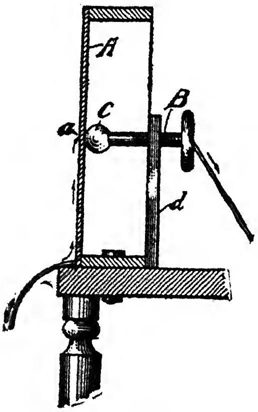


FIG. 17.—Berliner's transmitter.

The contributions of a few important inventions by Edison, Hughes and others should be recited before closing this brief outline of the history of the telephone during its infancy.

Bell's liquid transmitter embodies the main principle upon which subsequent successful battery transmitters are based —a battery furnishes the current and the transmitter actuated by the voice serves to modulate it. It was not long, however, before a much better means than Bell's was devised for putting this principle into practice.

The first step toward this was one made in 1877 by Emile Berliner, of Washington, D. C. He filed a caveat and later in the same year applied for a patent on a transmitter, depending upon the principle, already well known, that if the pressure between two conducting bodies forming part of an electric circuit be increased, the resistance of the path between them will be diminished, and, conversely, if the pressure between them be decreased, a corresponding increase of resistance will result.

Berliner's transmitter is shown in principle in Fig. 17, which is a reproduction of the main figure in his subsequently famous patent. In this A is a vibratory diaphragm of metal against

the center of which rests the metal ball *C* carried on a thumbscrew *B*. When the diaphragm is thrown into vibration by the voice, the pressure on the contact between it and the ball *C* is varied. This varies the resistance of the circuit through them and causes corresponding undulations in the current flowing.

At about this time Edison introduced carbon as the best material for the contacts to be used for varying the resistance with changes of pressure. This contribution by Edison was an important step in the development of telephony. Up to the present time no material has been found to even approach carbon for all-round effectiveness for the purpose. Of the 33,000,000 telephones in service in the world to-day, all, with perhaps insignificant exceptions, employ carbon transmitters.

An early type of Edison's carbon transmitter consisted simply of a button of compressed plumbago bearing against a small platinum disc secured to the diaphragm. The plumbago button was held against the diaphragm by a spring, the tension of which could be adjusted by a thumbscrew.

Many descriptions of Edison's carbon transmitters have alleged a peculiar property of carbon by which it changes its electrical resistance under changing pressure. Apparently however, carbon does not have this property in any such degree as would account for its variable resistance action in telephone transmitters. No marked change can be noticed in the resistance of carbon rods when subjected to a range of pressures from zero up to their crushing points. What is believed to account for the variable resistance effects in any carbon transmitter is the variation, under changing pressure, of the *intimacy of contact*, either between the surface of the carbon bodies themselves or between them and the contiguous electrodes, whether of carbon or metal. This, of course, does not explain why carbon is superior to all other conductors for variable resistance purposes, but it is, at least, in accordance with the loose-contact phenomenon, now to be referred to.

Professor David Edward Hughes made a most valuable contribution tending toward the perfection of the battery transmitter. By a series of interesting experiments, he demonstrated conclusively that a loose contact between the electrodes, no matter of what conducting substance they are composed, is better than a firm strong contact. The apparatus used in one of his earlier experiments, made in 1878, is shown in Fig. 18 and

consists simply of three wire nails, of which two form the terminals of the circuit containing a battery and a receiving instrument. The circuit was completed by a third nail laid loosely across the other two. Any vibrations in the air in the vicinity caused variations in the intimacy of contact between the nails and corresponding variations in the resistance of the circuit. This was a very inefficient form of transmitter, but it cleverly demonstrated the action of loose contact. Hughes became famous not only for his contributions to electrical science and practice but also for the simplicity and homely qualities of his laboratory apparatus—well exemplified in this experiment.

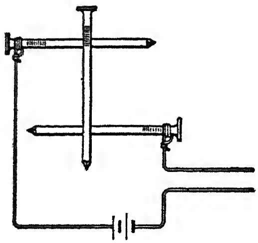


Fig. 18.—Hughes' nail microphone.

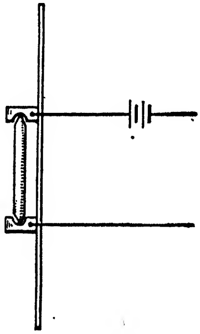


Fig. 19.—Hughes' carbon microphone.

Another form of transmitter devised by Hughes, and called by him the "microphone," is shown in Fig. 19. This consists of a small pencil of gas carbon, pointed at each end, supported between two blocks of carbon fastened to a diaphragm or sounding board. These blocks are hollowed out in such a manner as to hold the pencil loosely between them. The blocks form the terminals of the circuit. This instrument, though crude in form, is of remarkable delicacy and is well termed "microphone." The slightest noises in its vicinity, even those incapable of being heard by the ear alone, produce surprising effects in the receiving instrument. This particular form of instrument is, in fact, too delicate for ordinary use, as any jar or loud noise will cause the electrodes to break contact and produce deafening noises in the receiver. Practically all transmitters of to-day are of the loose-contact type, this having entirely superseded the first forms devised by Edison, in which the electrodes were more firmly held together.

In speaking of Professor Hughes' work on loose contacts and the microphone, an English electrical paper[1] says: "The microphone is a striking illustration of the truth that in science any phenomenon whatever may be turned to account. The trouble of one generation of scientists may be turned to the honor and service of the next. Electricians have long had sore reasons for regarding a 'bad contact' as an unmitigated nuisance, the instrument of the evil one, with no conceivable good in it, and no conceivable purpose except to annoy and tempt them into wickedness and an expression of hearty but ignominious emotion. Professor Hughes, however, has, with a wizard's power, transformed this electrician's bane into a professional glory and a public boon. Verily, there is a soul of virtue in things evil."

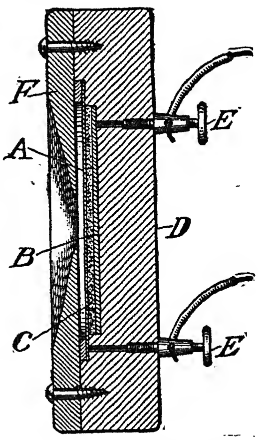


FIG. 20.—Hunnings' granular carbon transmitter.

In an article in *Nature*, June 27, 1878 Professor Hughes thus describes the conditions necessary for microphonic action: "If the pressure on the materials is not sufficient, we shall have a constant succession of interruptions of contact, and the galvanometer needle will indicate the fact. If the pressure on the materials is gradually increased the tones will be loud but wanting in distinctness, the galvanometer indicating interruptions; as the pressure is still increased, the tone becomes clearer, and the galvanometer will be stationary when a maximum of loudness and clearness is attained. If the pressure be further increased, the sounds become weaker, though very clear, and, as the pressure is still further augmented, the sounds die out (as if the speaker was talking and walking away at the same time) until a point is arrived at where there is complete silence."

Another valuable contribution was made in 1881 by Henry Hunnings, an English clergyman. He devised a transmitter wherein the variable resistance medium consisted of a mass of finely divided carbon granules held between two conducting plates. His transmitter is shown in Fig. 20. Between the metal diaphragm *A* and a parallel conducting plate *B*, both of which are securely mounted in a case formed by the block *D* and a

[1] *Telegraph Journal and Electrical Review*, July 1, 1878.

mouthpiece F, is a chamber filled with fine granules of carbon C. The diaphragm and plate form the terminals of the transmitter, and the current from the battery must, therefore, flow through the mass of granular carbon between. When the diaphragm is caused to vibrate by sound waves, its movements cause variations in the intimacy of contact between it and the granules, between the granules themselves and, to a less extent, between the granules and the rear plate. The resistance offered in the conducting path through them is thus varied and the desired current undulations produced. This transmitter, instead of having one or a few points of variable contact, has a multitude of them. It can carry a larger current without heating and, at the same time, produce greater changes in its resistance than the forms previously devised, and no ordinary sound can cause a total break between the electrodes. Although modern transmitters are quite different in form from this, their principle is the same. The use of granular carbon is well-nigh universal.

Edison made another important contribution to the early development of the telephone by using an induction coil in connection with the battery transmitter.

The induction coil used then and now for telephones employing a local battery is made as follows: Around a core formed of a bundle of soft iron wires is wound a few turns of comparatively heavy insulated copper wire. Outside of this is wound another coil consisting of a greater number of turns of fine insulated copper. The transmitter, together with the battery, is placed in a closed circuit with the winding of a few turns, while the winding of many turns is included directly in circuit with the line wire and the receiving instrument. This arrangement is shown diagrammatically in Fig. 21. A switch is included in the transmitter circuit to prevent the waste of battery while the telephone is not in use.

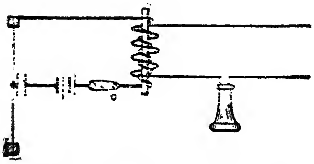


FIG. 21.—Transmitter in local circuit of induction coil.

The coarse winding is termed the "primary winding," because it is associated with the primary source of current, the battery; while the fine winding is termed the "secondary winding," because the currents flowing in it at the transmitting station are secondary, or induced, currents.

In action a current flowing in the primary winding produces a field of force extending through the core and into the surrounding space. Any changes caused by the transmitter in the strength of the current produce changes in the intensity of this field. As the secondary winding lies in this field, these changes will, by electromagnetic induction, cause currents to flow in the secondary winding and through the line wire to the distance receiving instrument. In good induction coils the electromotive forces set up in the secondary coil bear nearly the same ratio to the changes in electromotive force in the primary coil as the number of turns in the secondary bears to the number of turns in the primary.

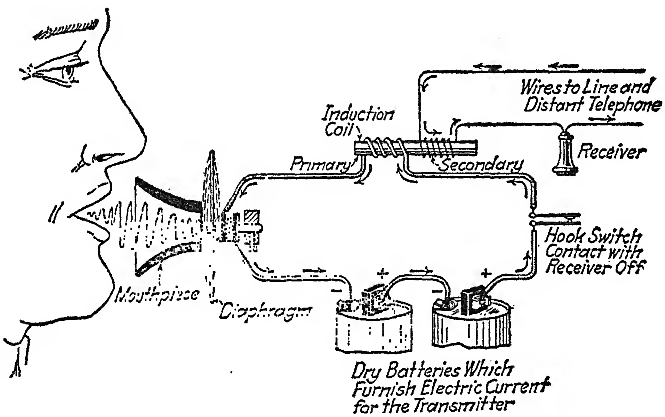


FIG. 22.—Talking circuit of a local battery telephone.

The use of the induction coil with the transmitter accomplishes two very important results: first, it enables the transmitter to operate in a circuit of comparatively low resistance, so that the changes in the resistance produced by the transmitter bear a larger ratio to the total resistance of the circuit and, consequently, produce greater changes in the current flowing; and, second, the step-up action of the coil produces higher electromotive force at the terminals of the transmitting station, thus enabling transmission to be effected over greater line resistances.

The arrangement diagrammatically shown in Fig. 21 is given somewhat pictorially in Fig. 22, this being one of Mr. Ray H. Manson's excellent pictorical diagrams. These two diagrams go no further than the actual talking apparatus. They show the "talking circuit" of a "local battery" telephone.

The difference between the two diagrams, both showing exactly the same circuit arrangement, deserves passing comment. The pictorial suggestiveness of Fig. 22, while of value to the beginner, is found to be unwarranted in practical work. In the first place, the illustration of parts with even this degree of detail would lead to impossible confusion in a really complex diagram. In the second place, such detail is unnecessary. Once the underlying principles of a piece of apparatus are understood, it may be represented by the merest symbol. To illustrate further, Fig. 23 shows exactly the same arrangement with even greater simplicity than that of Fig. 21. Thus has come about what may be called the "sign language of the telephone"—of which more will be said later.

We have now traced the early development of the elemental parts of the "talking set," the transmitter, receiver and induction

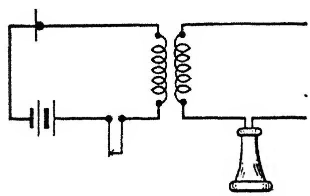


Fig. 23.—Talking circuit of local battery telephone.

coil—the parts actually used in taking the sounds of the speaker's voice and reproducing similar sounds in the listener's ear. Briefly reviewing this development:

Bell gave us the fundamental method involved in all telephony today. In addition, his receiver is the prototype of all modern commercial receivers. He also is responsible for the broad idea of the variable resistance transmitter.

Berliner contributed the idea of causing varying pressure between two transmitter electrodes to vary the resistance of the circuit; Edison introduced carbon as outstandingly the best electrode material; Hughes gave the knowledge of the loose-contact principle; Hunnings the use of granular carbon and, finally, Edison the induction coil as an adjunct to the transmitter.

All commercial telephones today employ the principles introduced by these men.

CHAPTER IV

THE TELEPHONE SYSTEM

To the average layman the *telephone* is merely an instrument, but to the telephone man it is a *system*. According to his experience and breadth of vision, the system he visualizes may be the small collection of associated apparatus with which he is immediately concerned; it may be the telephone exchange of his village, town or city; or it may be some grander conception, as of a universal system, which, by wire lanes and ether lanes, shall so cover the face of the globe as to place all people, whether on land or sea or in the air, in speaking relation with one another—the ultimate destiny of the art. Any one of these interpretations and many others may be correct. The term "telephone system" may comprehend little or much.

In the broader view, the most comprehensive telephone system so far actually achieved is that covering the entire United States from ocean to ocean and reaching into Canada, Mexico, South America and across the Atlantic to most of Europe. Some idea of the extent to which this covers the United States may be gained from Fig. 24, in which each dot represents an exchange area and each line a wire route.

The recent linking together of the United States and England by several separate wireless telephone channels has permitted the establishment of daily commercial service, which allows the people within reach of the vast wire networks of the United States and of Western Europe to converse with one another. A transatlantic telephone cable is in course of construction to make this service more reliable and to give more channels. Truly these are impressive steps toward the concept of a world-wide telephone system suggested in the opening paragraph. That concept is not as Utopian as it may at first appear. From the purely technical standpoint it could be achieved now. No doubt the telephone system of the world will be extended into all civilized countries as fast as it is commercially warranted.

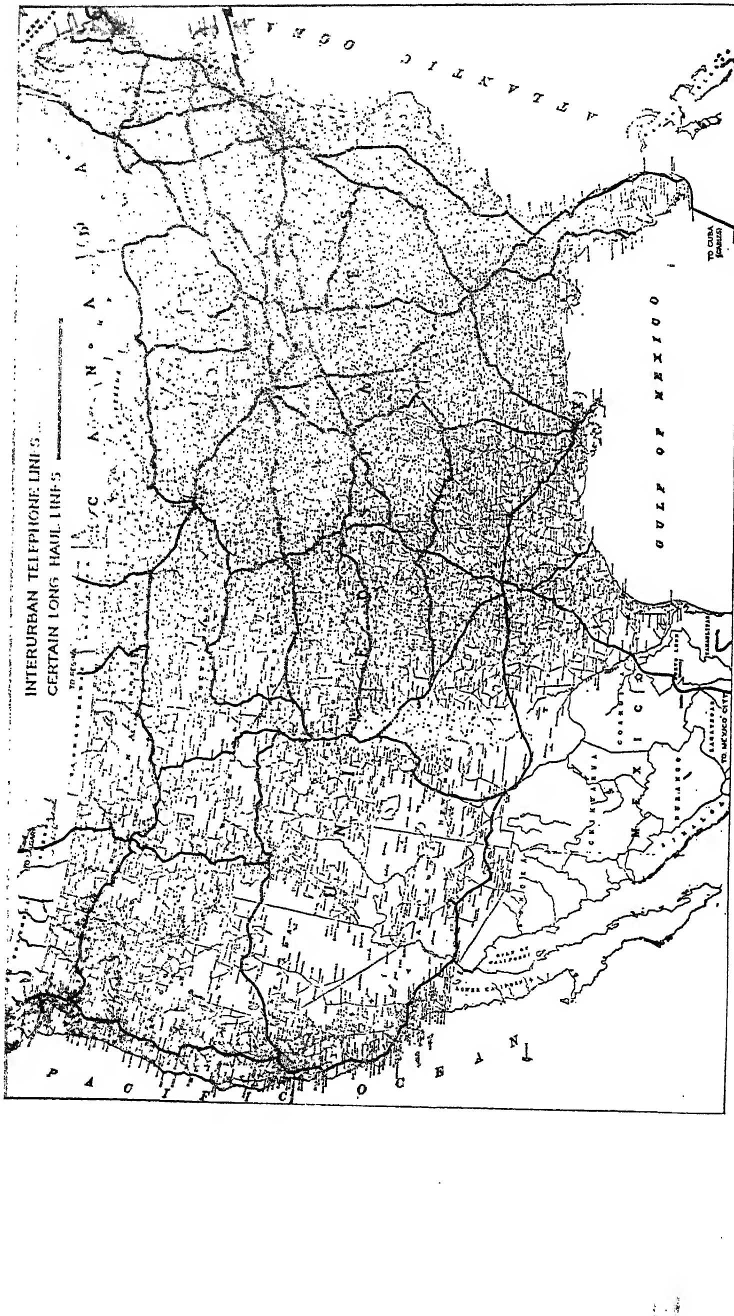


FIG. 24.—Toll line system in United States. *(Courtesy of Bell Telephone Laboratories.)*

But comprised within such a supertelephone system there are innumerable other systems, of many kinds and of varying degrees of complexity, performing numerous and diverse functions.

In the first place, there are the exchange systems themselves. Each of these is a complete system of local intercommunication. They are of many kinds and may be classified in various ways. For instance, there are manual, automatic and semiautomatic exchanges, according to whether the central office switching is done by hand, by machine, or by a combination of the two; and, on another basis of classification, there are single or multi-office exchanges, according to whether there are one or several central offices serving the exchange area in question.

Each exchange system in turn comprises within itself a number of minor systems, each performing a separate function or group of functions. Among these may be mentioned: systems by which current is supplied to the transmitters of the telephones; systems by which the telephone user signals to the central office operator; systems of attaching one or several telephones to a single line; systems of selectively or non-selectively ringing the bells on a line to which several telephones are connected; systems by which the patrons are charged for the service rendered, and so on.

Then, as distinguished from the exchange system, there is the long-distance system. No one has succeeded in drawing a sharp line to show where the exchange system ends and the long-distance system begins. As a matter of fact, there is no such real line, for the two are inextricably interwoven, and any such line of demarcation as is needed for administration purposes must, of necessity, be drawn in arbitrary fashion.

On the map the long-distance system looks rather simple, consisting merely of lines connecting the various dots representing the cities and towns. Actually, it is far more complex, for, in addition to the lines, there are long-distance switching systems and various systems for improving the efficiency of the lines. Among the latter are phantom, composite and carrier current systems by which each line is made to carry more than one message at a time. There are also the loading and repeater systems which allow small wires to serve where large ones would otherwise be required. Altogether, the long-distance system is one of the most remarkable triumphs of engineering. Some of its

devices, without which electrical communication on its present scale would be impossible, have come about only through the combination of most ingenious inventive skill, the deepest research in the realm of modern science and mathematical reasoning of a very high order.

Throughout the whole realm of telephony, the system must be kept in mind. Just as each individual piece of equipment must function properly within itself and also with respect to all the other pieces which go to make up its particular minor system, so all the minor systems must function properly within themselves and in relation to all the other minor systems which go to make up their particular major system. And, in the last analysis, all of the major systems must, with proper coordination, act as units in that comprehensive scheme which, for want of a better name, I have called the *supersystem.*

It is the purpose of this chapter to sketch in merest outline the course of development of some of the important phases of telephone systems. The details of this sketch will be filled in as far as possible when the various elements of apparatus, circuits, outside plant construction and method are taken up in the chapters to follow.

Such a preliminary outline should facilitate the more detailed discussion. It should also help one to keep in mind that the things being treated of are not detached units but are, in large measure, intimately correlated in function, and often in form, such other parts comprised within the system.

Talking Apparatus.—The germ of the telephone system was of course Bell's instrument. It was the thing out of which the whole telephone industry has grown. Let us first see how this germ *per se* has developed. In its original form it was a rudimentary affair. Its transmitting power was feeble, it was fragile, clumsy, unstable and, on the whole, quite unfit for commercial use even under the limited requirements of its time. Without considering the effects of improvements that have been made in telephone lines and in other parts of the system, some idea of the extent of the development of the instrument itself may be gained when it is stated that the telephone of to-day is of itself about thirty-two thousand times more powerful in its ability to transmit speech from person to person than was Bell's original instrument. Moreover, the present instrument is so rugged as to successfully stand rough handling and other vicis-

situdes of use, is pleasing in appearance, convenient in form, permanent of adjustment and, although complicated, is withal a thoroughly practical affair. In numbers, it has grown from one to many millions.

Signaling Apparatus.—The talking instrument alone would be of restricted utility. The early telephone had difficulty in making itself heard when applied directly to the ear and, unless too harshly treated, was quite unable to attract the attention of persons even a few feet away. Signaling apparatus had to be devised. The problem was not so easily solved as with the mechanical telephone, where, by means of mere thumps on the diaphragm at the transmitting station, sufficiently loud sound signals were produced by the diaphragm of the receiving station. The electric telephone was too delicate a device for such rough usage; but it is interesting to note that early attempts were made to signal in this way, as in the case of "Watson's thumper" later to be described.

It was natural that early efforts at signaling should have been by means of battery current and battery bells. This, however, involved difficulties with respect to maintaining batteries of sufficiently high voltage and bells of sufficient reliability. The use of both the battery and the battery bell for signaling purposes passed out of the art at an early date. Later, the battery came back, but not the battery bell.

To meet the early deficiency of the battery as a source of signaling current, the so-called "magneto generator" was adopted. At first this was a mere adaptation of the old magneto electric machine devised primarily for giving people electrical shocks. It was in fact nothing more than a small hand-operated alternating-current dynamo. As later developed, and as still used in rural communities, it had a coil of wire which by means of a hand crank and suitable gearing could be made to revolve rapidly between the poles of a permanent horseshoe magnet. This change from battery current to hand-operated generator current was apparently contrary to the trend of modern development, since it substituted man power for fuel power. Nevertheless, it was an important step, for it solved the problem of generating currents that were well suited for distant signaling. The alternating currents of the magneto generator had sufficient power and voltage to be effective over lines of considerable resistance, and the range of frequency easily attained was about

right for successfully operating signal bells. The availability of this kind of signaling current at once led to the production of a simpler and more rugged form of bell, the polarized bell or "ringer," as it is generally called. This bell was developed by Watson in the form essentially as used today.

The Telephone Set.—With the development of suitable signaling apparatus, the prototype of what we now call the "magneto telephone"[1] or "magneto telephone set" employed in "magneto systems" was evolved. To-day we speak of the transmitter, receiver and induction coil as forming the "talking set" and of the combination of generator and ringer as the "signaling set." As the talking and the signaling sets are used alternately, it was found necessary to provide, within the telephone, some sort of switching device so that they could be alternately connected with the line, and so that neither set would interfere with the functioning of the other. This led to the "hook-switch" an important element of all telephone sets. Because it is so made and placed as to form the only object upon which conveniently to hang the receiver after use, it automatically carries out this switching function without thought on the part of the telephone user.

The magneto telephone was used almost universally until a few years before the beginning of the present century. Then, as will be shown, there began a reversion to the use of battery as a source of signaling current. One of the results of this has been the doing away with the magneto generator except in the telephone systems of small communities, where the magneto system still survives. For such use, nothing as good has been found to supplant it. Watson's polarized bell, however, has survived in universal use. Like Bell's receiver, it will be found in practically every telephone station throughout the world to-day.

The magneto system required that each telephone station be provided with two independent sources of electrical energy; one a battery to supply direct current to the transmitter, the other the magneto generator to supply alternating current for

[1] The term "magneto telephone" has two different connotations. Ordinarily, in present-day parlance, it means a telephone set that is provided with a magneto generator for signaling. Frequently, however, it is used in its eariler sense as applying to Bell's original instrument or to its outgrowth, the modern receiver or hand telephone.

operating the distant signals. This was an uneconomical arrangement. It meant a great scattering both of investment and of maintenance effort. To-day, practically all except rural communities are served by what is called the "common battery system." In these both sources of energy at the telephone stations have been done away with, thus bringing about a great simplification of the telephone set. The reason for doing this and the manner of its accomplishment may be best understood in connection with the treatment of exchange systems.

The telephone set was thus reduced to its lowest terms, both in regard to the kind of things comprised within it and to the simplicity of its operation by the user.

Party Lines.—The simplest system of telephone communication is formed by two telephones permanently joined by a single line. While useful, the limitations are obvious. These limitations soon gave way under the avalanche of inventive effort that began with the introduction of the telephone, and that is still going on with ever increasing vigor.

The first efforts toward making the telephone system available for communication between more than two points were in the direction of connecting more than two telephones with a single line. Thus arose what we now call the "party line," meaning, of course, the "multiparty line." Such an isolated party line, even when extended to its utmost capacity in point of the number of stations connected, can, at best, serve for communication between a limited number of telephones and for only one conversation at a time. Notwithstanding these limitations, the isolated party line is still used, particularly in rural districts where the number of stations in a community is but few, or in other cases where a few stations of common interest are to be served. The isolated party line still seems the best if not the only practical solution for the intercommunication problem in certain pioneer rural districts, where the settlers are so few and their combined resources so meager as to make a more elaborate system unavailable. It has formed the nucleus about which many larger systems have been built.

There are cases in the western part of the United States where a single telephone line runs through a narrow mountain valley connecting the telephones of as many as forty or fifty farmers. This, of course, is not the best form of telephone service, but it is a veritable boon to those who, without it, would

often be almost completely isolated. From this sort of thing, the so-called "farmer's line" business has grown up. It is a troublesome and also important factor of telephone service; troublesome because of the difficulty of affording adequate service under adverse conditions at a price within the range of the patrons; and important because necessary as a part of the large conception of what telephone service should be.

The Exchange.—The idea of the telephone exchange, whereby a number of lines are brought to a common point and there provided with facilities for connecting and disconnecting them at will, opened the way to the complete removal of the limitations inherent in the isolated line.

As pointed out in an earlier work,[1] the first idea of an exchange that I have been able to find was that of Dumont in 1851.[2]

Not having the telephone, Dumont based his idea on the use of telegraph instruments for sending and receiving messages. As indicating the extent to which he had grasped the general idea of grouping lines and of connecting them for communication, one of the figures of his British patent is reproduced here as Fig. 25. The central point marked 0 in the diagram he called the "central station." This station and all of the subcenters numbered from 1 to 16 were provided with apparatus for connecting the lines. The mode of operation may be understood from the following quotation from Dumont's specification:

"Each house has an electric wire connecting it to the station, and a double electric apparatus, one at the subscriber's house and the other at the neighbouring station to which it is connected. The houses are telegraphically numbered from one to any number. Thus, suppose the house number 3, connected to station No. 1, and the occupant of house No. 3 is desirous to communicate with the house 287, connected with the station 12, the subscriber of the house No. 3 signalises to station 1, the number 287. The clerk of the station No. 1 then directs the central station 0 to place the wire 1.0 in communication with the wire 0.12. When this is done, the clerk signalises to station 12 the No. 287. The clerk at station 12 then connects wires 0.12 with the wire 12.287.

"This accomplished, the clerk of station No. 1 finally connects the wires 3 and 1.0 and a direct and intermediate communication

[1] "American Telephone Practice," 4th ed., p. 174, 1905.
[2] British Patent 13,497, 1851, to Dumont.

is then established between the houses 3 and 287, and both subscribers may correspond privately."

We must accord to Dumont not only the basic idea of exchange operation but also of multioffice operation with what we now

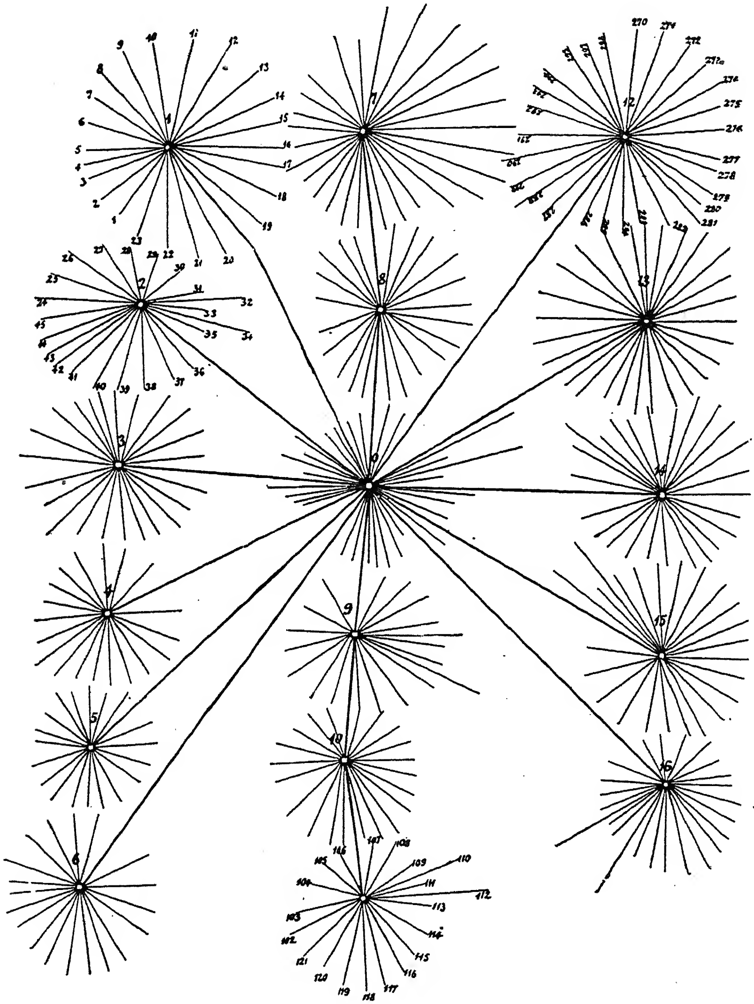


FIG. 25.—Dumont's telegraph exchange.

call "trunks" extending between the offices to enable the subscribers of one office to talk with those of another. Evidently he did not anticipate much traffic, for, instead of employing

"direct trunks" between each pair of offices as we now do in congested areas, he centered all trunks at one office and routed all calls through that office by connecting the trunks in tandem. This system of "tandem trunks," particularly as illustrated by the string of offices 7, 8, 0, 9, 10, etc., extending down the center of Fig. 25, is what we employ to-day in connecting widely separated and relatively unimportant offices, where the traffic does not warrant direct trunks between each pair of offices. Furthermore, Dumont is apparently the first to have designated the patron of such a service as the "subscriber."

So far as is known, nothing ever came of Dumont's idea. When the world was ready, even for a telegraph exchange, the invention apparently had to be made again. We think, however, that the name of Francois Marcelin Aristide Dumont, a French engineer, deserves a place in the telephonic hall of fame.

A so-called "telegraph exchange" was later established in London, evidently without the teaching of Dumont in mind, for this London exchange contained no equipment for interconnecting the lines. It merely permitted each subscriber to communicate with the central office.[1]

In 1874, just before the birth of the telephone, a telegraph exchange system was put into use in New York City. It was for connecting lawyers who chose to subscribe and was, for this reason, called the "law system." The telegraph instruments, which were of the dial pattern, were subsequently replaced by telephone instruments, and thus originated, as far as I am able to find, the first telephone exchange system. It was from this particular system that the well-known law telephone system, widely used in later years, took its name.

Apparently the telephone exchange had its practical beginnings in about 1878. These beginnings were made almost simultaneously in a number of cities in the United States. Few had any conception of its real significance. But what did the first telephone engineer think about it? Bell had the vision to see with remarkable clarity some of the possibilities of the thing which his genius had created. In a letter, dated March 25, 1878, written to British capitalists he says, in part:

"The simple and inexpensive nature of the telephone, on the other hand, renders it possible to connect every man's house, office, or manufactory with a central station, so as to give him

[1] KINGSBURY, J. E., "The Telephone and Telephone Exchanges," p. 80, Longmans, Green and Co., London, 1915.

the benefit of direct telephonic communication with his neighbors. at a cost not greater than that incurred by gas or water.

"At the present time, we have a perfect network of gas-pipes and water-pipes throughout our large cities. We have main pipes laid under the streets communicating by side pipes with the various dwellings, enabling the members to draw their supplies of gas and water from a common source.

"In a similar manner, it is conceivable that cables of telephone wires could be laid underground, or suspended overhead, communicating by branch wires with private dwellings, country houses, shops, manufactories, etc., etc., uniting them through the main cable with a central office where the wire could be connected as desired, establishing direct communication between any two places in the city. Such a plan as this, though impracticable at the present moment, will, I firmly believe, be the outcome of the introduction of the telephone to the public. Not only so, but I believe in the future wires will unite the head offices of the Telephone Company in different cities, and a man in one part of the country may communicate by word of mouth with another in a different place."

With the increasing realization of the possibilities of the telephone exchange, there began a period of intense development to provide the devices and the systems necessary to utilize them. But as the development progressed so did the possibilities. Each accomplishment opened up new visions of things yet to be done. And this process is still going on. In 1926, for instance, more than two new telephone patents were issued in the United States for every one that expired.

Telephone Switching.—The course of telephone exchange development may be illustrated most strikingly by considering the switching art, which relates to the appliances and methods employed at the central office by which lines are connected for conversation and afterward disconnected.

At the outset there were no switchboards except the very rudimentary ones that had been employed in telegraphy. These were at first used, but it soon became apparent that they could not cope with the more exacting and enormously extended requirements of telephone exchange work. The whole switching art had to be built up from practically nothing.

At first these switchboards were required to interconnect perhaps a dozen lines or less. Equipment had to be developed

even for this now simple requirement. The old telegraph switchboard of the crossbar type proved cumbersome and unreliable. In its place was developed a type wherein each line terminated at the central office in a connection socket or "jack" and in a visual signal or "annunciator." In order to connect the lines for conversation, flexible conducting cords were used terminating at each end in a metallic plug adapted to fit into the jacks. In order to assure firm contact between the plug and the jack, and also to accomplish certain switching operations, each jack was provided with a spring which was lifted by the inserted plug, hence the name "springjack."

The Manual Switchboard.—Thus the manual switchboard in telephony had its beginning. It is still used more extensively than any other type and its development has been one of the outstanding achievements of the art.

It must be remembered that in this development, the switchboard had at all times to be coordinated with the telephones at the subscribers' stations and *vice versa.* It is, therefore, natural that the first switchboards were of the so-called "magneto type," employing electromagnetic annunciators as line signals because these were well adapted to respond to the alternating currents from the magneto generators of the subscriber's telephones. These early switchboards were adapted to use on grounded lines only because, at that time, no one realized that any other kind of line would be required.

At first the number of lines was so small and the "traffic" so light that only one operator was necessary. Later as the number of lines and traffic increased the "load" had to be divided. This was done by assigning to each of several operators as many lines as she could conveniently attend and serve. The jacks and annunciators thus assigned to different operators were located on different sections or "positions" of the switchboard. As a result of this distribution, the telephone engineer was soon confronted with the problem of connecting two lines together whose jacks were not within reach of a single operator. This led to the so-called "transfer switchboard," employing local trunk lines extending between the sections or positions occupied by the different operators. By means of these, an operator answering a call on one of the lines assigned to her position could "transfer" the connection to the operator at whose position the

called line terminated, the connection between the two lines being established through the trunk.

This plan of employing intermediate trunks and two operators for completing a connection between two lines terminating in the same office was obviously uneconomical. A much better plan was soon devised. This resulted in the so-called "multiple switchboard" of which Leroy B. Firman, of Chicago, was the first proponent.

The idea back of the name multiple switchboard is that of multiple terminals for each line. The line signals and the corresponding incoming terminals ("answering jacks"), of which there is one for each line, are distributed in groups among the operators' positions in accordance with the ability of the operators to attend to them, exactly as in the transfer system. In addition, each line is provided with a number of outgoing terminals ("multiple jacks"), and these are so placed along the face of the board that one of them will lie within the field directly reached by each operator. By this means each operator, no matter what position she occupies, has within her reach a terminal or jack for every line. Without the assistance of any other operator, therefore, she may answer any call on the group of lines whose answering jacks are located on her position and complete the connection with any other line in the entire office. The number of lines for which this multiple arrangement is possible is evidently limited by the number of terminals that can be placed within the field of reach of an operator.

With the growth of the number of subscribers' stations and lines, the question of conservation of space on the face of the switchboard became increasingly serious. Also, the necessity of providing and maintaining a separate battery and magneto generator at each subscriber's station became more and more irksome. Relief in these and in other respects was afforded by the advent, in the early nineties, of the "common-battery" switchboard. This was a revolutionary change and has been universally adopted in all exchanges except those serving small communities.

The common-battery system has also been called the "central-energy" system. In it, as each of these names indicates, all the hitherto scattered sources of electrical energy are replaced by a common source located at the central office. A single battery of sufficient voltage and capacity is used to supply talking current

to all of the subscribers' transmitters. Having this current supply for transmitter operation, it is equally available for operating the central office signals. Hence, as a result of this one change, the necessity for the magneto generator and the local battery at the subscriber's station disappeared.

The economy of this step was obviously very great. Instead of having perhaps ten thousand local batteries and ten thousand magneto generators located under ten thousand different roofs, there was one large battery located under one roof. Economies were also permitted in other directions. Among them was that of space on the face of the switchboard. The steady voltage common battery lent itself admirably to the use of incandescent lamps for central office signals, whereas the magneto generators did not. As a result, miniature incandescent lamps replaced the old magnetic drops on the switchboard. They occupied but a fraction of the space, gave better signals, were automatically self-effacing and were more reliable in operation.

With the development of the common-battery multiple switchboard, manual switching reached the basis on which it rests to-day. The practical ultimate capacity of a single multiple switchboard unit has been found to be in the neighborhood of 10,000 lines. This, therefore, is the limiting capacity of a single office. When more lines than that are to be served, other offices are added, and the old transfer idea again resorted to. In some of the more congested metropolitan areas, the density of telephone development is so great as to require several 10,000-line switchboard units in a single building serving a relatively small area.

The Automatic Switchboard.—It is surprising that almost as soon as the idea of central office switching was put into practice, a few ambitious inventors conceived of doing this switching automatically. They proposed to dispense with the operator and, instead, employ means whereby the subscriber could, by remote control, perform his own switching operations at the central office.

The efforts toward such a difficult accomplishment while the art was yet so young were thought to be misdirected. So-called "practical" men had nothing but intolerance toward the idea, and ridicule was heaped upon those who persisted in such efforts. For many years little headway was made. This is not surprising, because the technique of the telephone art, even in its

simpler phases, had not yet been sufficiently developed to form a foundation for the practical accomplishment of this difficult feat.

These earlier workers in the automatic field were followed by such men as Alexander E. Keith of Chicago, Illinois, George William Lorimer of Brantford, Ontario, and later of Ohio, and a host of others who had not only the vision but also the practical ability to make their dreams come true. A few automatic exchanges that may be considered practical were established in the United States and Canada during the early years of the present century. Among the more notable may be mentioned those at Fall River, Massachusetts, in 1903, Grand Rapids, Michigan, in 1904, Columbus, Ohio, and Peterboro, Ontario, in 1905. The number of such exchanges grew, and gradually the idea began to dawn on those in control of the larger affairs of the telephone business that the automatic exchange was a really practical affair.

It was not until the World War, however, that the real economic possibilities of automatic or machine switching came to be generally appreciated. As a result of the conditions during and following the war, the market value of female labor was greatly increased. Young women, who previously had been satisfied with relatively low wages for telephone operating, found more remunerative if not more attractive work in other fields. A readjustment of operator wages had of necessity to follow. This was one factor which greatly altered the relative economies of the manual and the machine methods of operating. That the ability of the manual system to cope with the constantly increasing traffic loads was being more and more severely taxed in large metropolitan areas was another factor. More will be said later of the relative economies of the two systems, but at this point it may be stated that there is general belief among telephone engineers and those companies and governments who are responsible for telephone operation in a large way, that the machine-switching system in some of its forms will gradually replace the manual in all large cities. How far down in the scale of size of communities this change will go is still a matter of wide difference of opinion and consequently of much discussion.

There has also been much discussion as to how the machine-switching equipment at the central office should be controlled, or, more strictly speaking, by whom it should be controlled.

The plan now generally adopted is to have the subscribers do it. Another plan, whose advocates are decidedly in the minority, is to have a comparatively few operators do it. It is simply a question of whether substantially the same switching machines at the central office shall be directed by the dials on the subscriber's telephones or by "adding machine key boards" manipulated by operators, who receive their instructions from the subscribers. There is much to be said for each plan.

For reasons not altogether logical, the subscriber-controlled systems are referred to by such names as "full automatic" or "full mechanical," and the operator-controlled systems are variously designated as "semiautomatic," "semimechanical" and "automanual." In any event, they are all properly "machine-switching systems."

The Telephone Line.—In thus briefly tracing the development of the telephone *per se* and of central office switching practice, it must be kept in mind that no less radical and no less important developments have been simultaneously going on with respect to the lines that connect the subscribers with the central office and also with those that connect distant central offices with each other.

In the early days the fact, previously discovered by Steinheil, that the earth could be used for the "return" conductor of an electric circuit was made use of in telephony, as it already had been, and still is, in telegraphy. Accordingly, a single wire was used for the line and this was connected to earth through the terminal instruments at each end to effect the return path. Such lines are termed "grounded lines."

At first outdoor telephone lines were made of iron. Ordinary copper wire was recognized as being a far better conductor and also as being non-corrosive under weather conditions, but it had too little tensile strength to stand the stresses encountered in open-wire construction. An important step was taken by Thomas B. Doolittle, who, as early as 1877, experimented in the direction of making a sufficiently strong wire of copper by the simple expedient of omitting the annealings between the successive drawing operations. Hard-drawn copper wire, resulting from Doolittle's investigations, has become the accepted standard for aerial telephone line wire, except in the case of relatively short and less important lines where iron, on account of its cheapness, is still used to a considerable extent.

The grounded line almost proved the undoing of the telephone business in its early days. With it strange and unaccountable noises were heard in the telephone instruments, some of which even to-day are only imperfectly understood. Some of these noises were due to natural phenomena, others to "man-made" causes. Of the former, induced currents from distant lightning and earth currents, possibly due to changes in the earth's magnetic field, may be mentioned. Of the latter, induction and leakage currents from telegraph or other telephone lines were at first the principal if not the only ones. But these were soon supplemented by the far more devastating effects of stray and induced currents from the electric railway, to say nothing of the induction from electric light and power distribution systems.

The sole remedy for all of this trouble from external sources was found to be in the use of the "metallic circuit." This means a circuit formed of two wires of equal size and the same material placed as close together as practicable and occupying, as far as possible, equal average distances from all external disturbing sources.

In spite of the increasing difficulties the general use of grounded lines persisted for more than a decade, partly because of the obvious economy of using one wire instead of two and partly because of uncertainty as to the real remedy. The coming of the electric railway in the late eighties precipitated the issue. Litigation with the railway companies ensued in which the telephone interests finally lost.

Regardless of the legal aspects, it is now easy to see that the inevitable thing for the telephone companies to do was to adopt the metallic circuit. It is now almost universally used. Only in a comparatively few locations that are remote from other interfering electrical systems, and in those cases where the necessity for economy forces the user to tolerate the disturbance does the grounded line still persist.

The elimination of line disturbances was not the only problem with respect to telephone lines. Another was that of wire congestion. Practically all outside line wires were suspended from housetop fixtures, chimneys and poles. There was no system to it. The wires ran helter-skelter with slight attempt at orderly arrangement. The congestion and the confusion constantly increased with the growing number of telephones, and the adoption of metallic circuits, while curing one trouble, made this one worse.

The tangle of overhead wires in the downtown areas of cities became a menace to life and property, to say nothing of its bad appearance. To clean up this mess and to provide for the great expansion to come was a Herculean task. There was no experience to serve as guide, for never before had the necessity arisen of placing so many wires in such close proximity with each other. The development of the technique for the orderly and effective arrangement of bare wires on poles and for the placing of a large number of wires in cables and disposing these overhead, underground or underwater has been one of the ever present problems of telephony. Its various phases cannot be even referred to in the outline sketch of this chapter; but to illustrate the nature of some of the problems involved and of some of the results achieved, the development of one item alone, telephone cable, will be briefly mentioned.

Telephone Cable.—The use of bare wires requires a considerable separation of the wires from each other. One reason for this is the necessity of preventing actual contact when they are violently agitated by wind or other causes. This separation that is a requisite of "open-wire construction" is at variance with the space economy that is essential when it becomes necessary to stow away a large number of wires in a very small space. Out of such considerations arose the demand for telephone cable, wherein a large number of wires, carefully insulated from each other, are bound closely together under a common covering.

This had been done to a limited extent in telegraphy, but when the single-wire telephone lines were thus treated, it was found that their talking efficiency was impaired or ruined, and also that whatever was said on one wire could be heard almost equally well on all the others, a phenomenon known as "cross-talk." These facts alone would have sounded the death knell for the general use of grounded circuits even in the absence of other considerations.

At first it was thought that the remedy for cross-talk and other inductive disturbances in cable was to be found in the character of the insulation, but finally it was learned that the only panacea was to use two-wire circuits throughout, the wires being twisted together in such a way as to produce a complete transposition every few inches.

Naturally, rubber, on account of its high insulating and moisture resisting properties, was one of the early materials

used for insulating the cable wires. This proved objectionable because of the high specific inductive capacity of rubber and its high cost. Obviously, because of its low specific inductive capacity, air would be the ideal insulator, but it does not afford the mechanical properties necessary to keep the wires apart. A loose wrapping of dry paper was found to be the closest approach to this ideal. It afforded the necessary mechanical separation, was of itself a good insulator, and its many interstices contained a large amount of air. Dry paper, however, is exceedingly susceptible to moisture and moisture is ruinous to good insulation. This difficulty was overcome by enclosing the bundle of wires in a continuous lead pipe forming an air-tight sheath. Such is the cable now universally used for the wires of telephone lines.

Early cables had a maximum capacity of about 50 pairs of wires. Later, 200-pair, 400-pair and 600-pair cables were successively evolved, the limiting outside diameter remaining approximately the same—somewhat less than 3 inches. At this stage, the increase in size stopped for about a decade, but the progress toward a larger number of pairs was again resumed, and we now have 1,800-pair cables containing over 3,600 wires. Cables of even larger capacities are in sight.

The economies resulting from this cable development have been enormous. They are due not only to the saving in copper and other material but also to other factors, such as the saving in poles or underground ducts. Some idea of what the use of this type of cable has meant in the telephone business may be gained from the fact that a single one of these cables, less than 3 inches in diameter, may be made to carry as many wires as, with ordinary open-wire spacing, would be carried on thirty parallel pole lines each having twelve 10-pin crossarms.

The constantly increasing demand for cables carrying more and more wires brought other problems than those involved in making, putting in place and maintaining the cables themselves. The maximum outside diameter of the cables has remained the same, being limited by the sizes of existing underground ducts. On this account an increased number of wires has meant smaller wires with consequently increased electrical resistance; and more crowding together of the wires has meant greater electrostatic capacity between the conductors of a pair. These changes were in themselves deleterious to the transmission efficiency of the

conductors both for speech and signals. In order to maintain the required over all standard of transmission for long-distance as well as local communication, more powerful transmitting devices, more sensitive receiving devices and more efficient auxiliary devices were necessary, all working toward a system of balanced economy. Conversely, each improvement in transmission equipment has at once been met with the demand for cables with a greater number of conductors, the improved transmitting efficiency of the apparatus permitting a further sacrifice in conductor efficiency without overstepping the limiting requirements of satisfactory telephonic transmission. Nothing could better illustrate the interrelationship between the various parts of a telephone system.

So far this outline has considered the telephone line merely in so far as it affords a conducting path, in the ordinary sense, from one place to another. There remains to consider the very important work that has been done toward improving the *effectiveness* of the line. These efforts have taken three directions:

a. To make the line a better path for telephone currents. This has led to the practice called "loading."

b. To reinforce the energy supplied to the line at various points along its length. This has called forth the "telephone repeater."

c. To make a given line carry more than one message at a time. From this has grown "phantom," "simplex" and "composite" circuits and, finally, "carrier-current" systems.

Loading.—Every telephone line has a certain electrostatic capacity. The two wires may be considered as forming the plates of a condenser, and the insulation between them the dielectric. This electrostatic capacity instead of being localized, as in an ordinary condenser, is distributed throughout the length of the line. It will be shown that this distributed capacitance has a deleterious effect on the transmission of rapidly varying currents, tending not only to attenuate but also to distort voice currents. It will be shown that inductance coils, likewise, have a deleterious effect on the passage of such currents through them; and that, in a general way, inductance and capacitance act oppositely in producing these bad effects.

Oliver Heaviside, the great British mathematician and physicist, suggested that these opposing effects might be made to neutralize each other in the transmission of rapidly varying

currents. It was not until 1899, however, that Michael Idvorsky Pupin showed exactly how the distributed capacitance of a telephone line could be neutralized for a given range of frequencies by the employment of inductances of specific amounts located at stated intervals along the line. Pupin showed the way as a mathematical and physical conception. His work was taken up by Campbell and other mathematicians, physicists and engineers of the American Telephone and Telegraph Company, who have developed the art of "loading" to its present high state of efficiency.

The economies resulting from loading have been enormous. The use of loading coils permits very much smaller copper wires to be employed in the transmission of speech than would otherwise be required. This alone has had a profound influence in making possible the extensive use of the small-wire cable. The resulting savings are due not alone to the economy in copper but also to many collateral advantages, such as allowing the use of fewer and lighter poles and of a smaller number of underground ducts.

Repeaters.—Unlike loading, repeaters do not aim to improve the conducting path for the voice currents, but rather to reinforce these currents by introducing new sources of energy at various points along the line. In this way the amount of energy arriving at the distant point may be made substantially the same or even greater than that leaving the starting point.

The repeater has always been an attractive mark for inventors. Efforts to produce one naturally took the form of a mechanical coupling between the vibratory part of a receiver and that of a transmitter. In this way the receiver would, as it were, talk to the transmitter, and the transmitter, as usual, would act as a valve in causing modulations in the stronger current of a local source. It was easy to make these mechanical repeaters work in one direction. It was more difficult to make them work successfully in both directions. Moreover, the mechanical inertia of their vibratory systems was always a serious drawback, particularly since it acted unequally on vibrations of different frequencies, thus producing distortion.

In spite of these difficulties, Mr. Herbert M. Shreeve, of the American Telephone and Telegraph Company, produced a successful mechanical repeater which was used with moderate success on the transcontinental lines of the American Telephone

and Telegraph Company. It probably would have been further improved had not the vacuum tube repeater been produced at about that time. This, being devoid of mechanical inertia, was so vastly better than the mechanical repeater, both in principle and in practice, that it at once removed further incentive to work along the mechanical line. With the vacuum tube repeater practically distortionless amplification was possible, while with the mechanical repeater it was not.

To show what the repeater means on the New York-San Francisco line, the following quotation from a recent paper by Messers. H. H. Nance and O. B. Jacobs is given:[1]

"Without telephone repeaters but with other parts of a long telephone circuit unchanged, the delivery at the receiving end of the amount of power ordinarily obtained would require startlingly large amounts of power at other points in the circuit. For example, in the case of a San Francisco-New York connection, the amount of power ordinarily applied at San Francisco would be required near Harrisburg, Pa. All of the power introduced ordinarily at all points in the line would be required at a point near Pittsburgh. Power sufficient to light two 20-c.p. incandescent lamps would be necessary near Chicago, while the power of a five-kw. radio station would be required near Omaha. The requirements continue to rise rapidly until, near Rawlins, Wyo., a 50,000-kw. generator would have to deliver its entire rated capacity to the circuit, while at San Francisco, something in the order of the estimated *total world production of mechanical and electrical power would be needed.*

Let us suppose, however, that a 50,000-kw. generator delivered its entire output to the circuit at San Francisco, and overlook, for the moment, what would happen to the line if any such amount of energy were applied. The power received at New York would be of the order of one five-hundredth of a microwatt, *which would have to flow for about* 25,000 *years in order to equal the energy required to light a 25-watt lamp for one minute.*" (Italics mine.)

Simultaneous Messages.—Efforts to send more than one message at a time over a wire are older than the art of telephony. Bell was trying to do this telegraphically before he invented

[1] Nance and Jacobs, Transmission Features of Transcontinental Telephony, presented at the Pacific Coast Convention of the American Institute of Electrical Engineers, Salt Lake City, Utah, September 9, 1926.

the telephone. By properly combining two metallic circuit telephone lines a third or "phantom" circuit is made available, thus permitting three simultaneous conversations over two "physical" lines. Then there are various "simplex" and "composite" systems, wherein the two telephone wires of a line may, in addition to carrying telephone messages in the regular way, also carry one or several simultaneous telegraphic messages. These systems have long been in regular commercial use.

Of later origin, however, are the so-called "carrier-current" systems, by which a telephone line, used in all the foregoing ways for transmitting telephone and telegraph messages, may be made to carry as many as six additional telephone messages without interference. These carrier-current systems make use of the fact that vibrations having frequencies above about twenty thousand cycles per second are inaudible. Currents lying within successive bands of frequencies above the audible range, therefore, may be applied to the telephone line without interfering with its ordinary operations. Each of these successive bands of inaudible frequencies is made to afford a separate "channel" of telephonic communication. This lies at the very threshold of the radio art.

Some idea of the benefits derived from these methods of simultaneous transmission may be gained from the fact that on some of the main open-wire circuits of the transcontinental line as many as twenty two-way communication channels are derived from two pairs of wires. Of these, six are telephone and fourteen are telegraph channels.[1] Again, on a single conductor submarine cable, extending from Catalina Island to the mainland of California, eight different channels are secured, one an ordinary direct telegraph circuit, one an ordinary telephone circuit, and six carrier-current circuits.[2]

Interrelationships.—Telephone plant is comprised roughly within three general classifications, subscribers' station equipment, central offices and lines. In the foregoing pages something has been said of the trend of development of each, and of how these developments have necessarily been correlated in order to

[1] Nance and Jacobs, *op. cit.*

[2] Carrier Current on Submarine Cables, by H. W. Hitchcock, presented at the Pacific Coast Convention of the American Institute of Electrical Engineers, Salt Lake City, Utah, September 9, 1926.

produce not only an operative system but also a system properly balanced from the economic standpoint.

In the last analysis the telephone system includes *personnel* as well as *plant*. Regardless of the extent to which labor-saving machinery may be employed, the human element will always be of vast importance in it. Human effort must always be concerned not only with the workings of the system within itself but also in establishing and maintaining the proper relationships with the public served. And in this connection let it be kept in mind that the contact between the telephone system and the public it serves is one of peculiar intimacy. The telephone has a personal touch with its patrons that is of a different kind from that found in any other public utility service.

There is also a peculiar relationship as to both equipment and method not only with respect to the practices of a single company within itself but also to those of different companies whose lines require to be connected for long-distance service. The underlying reason for this is that the commodity supplied is *intercommunication*. Always two persons are involved in a telephone conversation and, frequently, two or more operating companies. A bad instrument or a faulty line at one subscriber's station, or an improper operating method on the part of his company, may affect the service of another subscriber, whether a neighbor or one located in the next county or state or even across the continent. No such interrelationship exists in like degree in any of the other public utilities. In the case of water, gas or electricity a consumer merely gets his commodity from a pipe or wire and, except for this, is largely unrelated to and independent of the other consumers.

If each operating company or other operating organization went its own way without regard to others, there would result a heterogeneous conglomeration of equipment and practices which, while sufficing for service within individual areas, would fall far short of the best possibilities of the industry. In following such a course, no company could possibly keep step with the progress of the art.

This correlation in order to be effective must be extended to all parts of the system. The ability to talk successfully between New York and San Francisco on overhead wires or between Boston and Washington on underground wires does not rest alone on the power or the sensitiveness of the telephone instruments at

the terminal stations. The most powerful and sensitive instruments known could not do it unless all other parts of the system were properly coordinated. Not only the subscribers' instruments themselves but also the relatively short subscribers' lines connecting the two stations to the respective central offices, the equipment at the central office by which the subscribers' lines are connected with the through trunk lines, the trunk lines themselves and, in fact, every link and element involved in the connection must be properly related to the others, in order that the limiting losses permitted by practical telephone transmission may not be exceeded. Somewhat less efficient transmitters, for instance, might suffice, but the sacrifice in this respect might require the expenditure of many millions of dollars in line conductors. And so with each element of the plant; undue loss in one calls for increased cost in others.

PART II

ELEMENTARY THEORY

This part attempts to lay a foundation of elementary theory upon which to rest an understanding of the processes and things involved in telephony. The science of sound in both its physical and physiological aspects is first dealt with to develop the fundamental requirements of speech reproduction. Then follows a similar discussion of the multi-frequency alternating currents which simulate the sound waves in composition and which are necessary in their reproduction at distant points. Thermionic emission is taken up and the principal uses of vacuum tubes in telephony are outlined. A final chapter reviews the principles of ferro-magnetism and discusses some of the new magnetic materials that are particularly applicable to the uses of telephony.

CHAPTER V

THE VIBRATIONS OF SOUND

The word "sound" as used in acoustics has two closely related but distinct meanings; one introspective, physiological and psychological and the other subjective and physical—one relates to a sensation, the other to the cause of it.

In the first sense sound is something entirely within us. It is the *sensation* occuring in the organs of hearing.

In the second sense sound is something that comes to us from without. It is purely physical and may be defined as a *disturbance* in the air or other media consisting of a succession of alternate waves of compression and rarefaction that tend to travel outwardly in all directions from the sounding body.

Under the first or physiological meaning there would be no sound if the disturbance in air was of such nature as not to be perceived by the ear; neither would there be sound in the case of an explosion with no ear within range to hear it. Under the second or physical meaning any periodic vibrations of air would constitute sound, whether of such character as to be perceived by the ear or not, or whether within reach of an ear or not. Under this second meaning, therefore, we have both audible and inaudible sounds.

Often, in an effort to avoid confusion between these two meanings of the same word, an attempt is made to use one of them to the exclusion of the other. Such attempts seem to lead inevitably to trouble, as when a writer, after defining sound as the "sensation perceived by the ear," proceeds in the next paragraph to speak of the "speed at which sound travels through air." The fact is that both of these meanings are well established in our language, and the context usually will leave no doubt as to which is intended.

In telephony we are vitally concerned with both phases of sound definition. From the physiological and psychological aspect we are interested in the capability and the behavior of the ear and the brain in perceiving and in interpreting vibrations of different characteristics. From the physical aspect

we are concerned with the study of the sound waves themselves. It is with this latter aspect that the present chapter will particularly deal, the consideration of the sensation of sound being specifically treated in the chapter next following.

It is the function of the telephone to transmit sounds to points distant from their source. Strictly speaking, this definition is not quite accurate. With the electric telephone the sound waves are not actually transmitted over the wires or other media, but, instead, the process is one where the energy of the originating sounds is transformed into electrical energy, which, in turn, is transmitted to the desired destination, and there reconverted into sounds more or less closely resembling the original. It would be more correct to say that it is the function of the telephone to reproduce sounds at points distant from their source. However, this somewhat loose use of the word "transmit," as involved in such expressions as "the electrical transmission of speech," has become a fixture in the language of the telephone. While less exact it is more descriptive.

Two branches of physics are primarily involved in the electrical transmission of speech and music—those of sound and electricity. There has been no lack of appreciation of the fact that the problem of telephony is largely an electrical problem, but, until recently, there seems to have been some failure to appreciate that back of that it is fundamentally a problem of acoustics, the science of sound.

We have long accepted Alexander Graham Bell's statement, that it was his greater knowledge of sound than of electricity that led him to invent the telephone, but, until recently, we have let the obvious lesson contained in that statement go unheeded. Bell did not solve all of telephony's problems of acoustics any more than he did all its problems of electricity, mechanics or business; yet for the most part, telephone workers have devoted their attention to these other things with too little regard for the science of sound.

Fortunately this condition is now rapidly being remedied. In addition to the fine scientific investigations which have been conducted from time to time largely by institutions of learning, there has been in recent years an intensive attack by telephone people on the problems of acoustics, with particular attention to the art of sound reproduction. The results have already been striking. Recently the phonograph has been revolutionized,

and great improvements have been made possible in telephone communication. The improvements in telephony have resulted not only in better quality and a greatly extended range of transmission but also in large economies as well.

The waves emanating from a sounding body tend to progress outwardly in all directions from the source. In a uniform medium, such as still air, the wave front is in the form of a constantly expanding sphere with the disturbing point at its center. While the wave front tends to move outwardly in all directions like the surface of a toy balloon that is being inflated, we must avoid the idea that the air itself actually has any such progressive movement. It is only the wave that progresses outwardly, the air particles themselves merely partaking of a small to-and-fro motion in directions coinciding with the radii of the sphere.

This idea of small longitudinal vibrations of the air particles resulting in a series of progressive waves is sometimes rather difficult to grasp. In his characteristic manner, Tyndall, in giving first ideas concerning longitudinal wave propagation, used the following quaint illustration in the first of his classic lectures on sound:[1]

"The possession of clear fundamental ideas is so important, that I propose to illustrate the propagation of sound by another homely but useful experiment. I have here five young assistants, A, B, C, D, and E, placed in a row, one behind the other, each boy's hands resting against the back of the boy in front of him. E is now foremost, and A finishes the row behind. I suddenly push A, A pushes B, and regains his upright position; B pushes C; C pushes D; D pushes E; each boy, after the transmission of the push, becoming himself erect. E, having nobody in front, is thrown forward. Had he been standing on the edge of a precipice, he would have fallen over; had he stood in contact with a window, he would have broken the glass; had he been close to a drum-head, he would have shaken the drum. We could thus transmit a push through a row of a hundred boys, each particular boy, however, only swaying to and fro. Thus, also, we send sound through the air, and shake the drum of a distant ear, while each particular particle of the air concerned in the transmission of the pulse makes only a small oscillation.

"But we have not yet extracted from our row of boys all that they can teach us. When A is pushed he may yield languidly,

[1] Tyndall, John, "Sound," D. Appleton & Company.

and thus tardily deliver up the motion to his neighbour B. B may do the same to C, C to D, D to E. In this way the motion might be transmitted with comparative slowness along the line. But A, when pushed, may, by a sharp muscular effort and sudden recoil, deliver up promptly his motion to B, and come himself to rest; B may do the same to C, C to D, and D to E, the motion being thus transmitted rapidly along the line. Now, this sharp muscular effort and sudden recoil is analogous to the *elasticity* of the air in the case of sound. In a wave of sound, a lamina of air, when urged against its neighbour lamina, delivers up its motion and recoils; and the more rapid this delivery and recoil, or in other words the greater the elasticity of the air, the greater is the velocity of the sound."

The arrangement of balls and springs suggested in Fig. 26 is of further help toward a conception of the way in which a pulse travels through an elastic medium. Each of the solid balls of the row is supported on a resilient rod, every ball being separated from its neighbor by a coiled spring. When the left-hand ball is forcibly moved to the right by a blow of the hammer, it flexes its supporting spring and compresses the spring at its right. This compression in the first coiled spring will cause a movement of the second ball toward the right, which, in turn, will compress the second coiled spring, and so on, a wave of compression going down the line of balls toward the right until, finally, the last ball is moved and taps the finger of the hand waiting to receive the sensation.

Here the elasticity of the springs corresponds to that of the air, and it is clear that the stiffer the springs (the greater their elasticity) the faster will be the progress of the impulse. Also the mass of the balls corresponds to the density of the air, and, obviously, the heavier the balls the slower will be the progress of the impulse. In such an arrangement, with fairly heavy balls and springs that are not too stiff, the progress of the impulse may be followed with the eye.

Again, after exploding a small gas balloon, Tyndall called to the attention of the audience the fact that they had all heard the sound, but that no particle of the air from the vicinity of the balloon had reached the ear of any person present. Concerning this he says:

"Scientific education ought to teach us to see the invisible as well as the visible in nature; to picture with the vision of the mind

those operations which entirely elude bodily vision. With regard to the point now under consideration, we must endeavour to form a definite image of a wave of sound. We must be able to see mentally the air particles when urged outwards by the explosion of our balloon crowding closely together; the particles immediately behind this condensation being separated more widely apart. We must, in short, be able to seize the conception that a sonorous wave consists of two portions, in the one of which

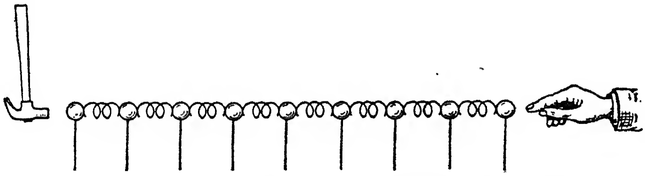


FIG. 26.—Transmission of impulse through elastic medium.

the air is more dense, and in the other of which it is less dense than usual. A condensation and rarefaction, then are the two constituents of a wave of sound."

So far we have dealt apparently with but a single disturbing impulse—the push given to one of Tyndall's "young assistants," the "pop" of his tiny balloon or the blow of the hammer of Fig. 26. Experience teaches us, however, that when such single disturbances occur in an elastic medium they produce not one wave but a series of diminishing waves which either rapidly or slowly die away, in much the same way that a pebble thrown into

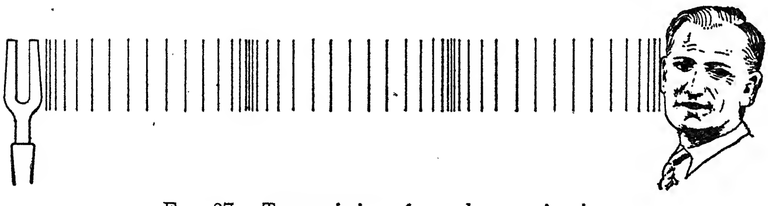


FIG. 27.—Transmission of sound waves in air.

the water produces not one wave but a train of them, each smaller than the one before it. In the case of the "pop" caused by the exploding balloon, the train of waves died out so rapidly as to leave only a single sudden sensation upon the ear.

As an example of a wave train that does not die out so suddenly but is maintained for a considerable period, because the sounding body continues in action, we may take the tuning fork shown in Fig. 27. When the fork is set in motion it continues its vibra-

tions, and thus creates a succession of waves which pass through the air to the ear. The alternating waves of compression and rarefaction are indicated in this figure by the varying degrees of closeness of the vertical lines intervening between the fork and the ear.

In the case of this long continued series of waves the ear would sense what we call a "tone"; and here we may observe a distinction in terminology. The sound of the exploding balloon was a "noise," that of the tuning fork a "tone," a distinction that will be referred to later.

That it takes sound an appreciable time to travel over considerable distances is a matter of commonplace observation. The time elapsing between a flash of distant lightning and the resulting sound of thunder is perhaps the most familiar example; another is that the steam from the whistle of a locomotive half a mile away may be seen some seconds before the sound is heard. The lightning flash and the sight of the steam reach us at the enormous speed of 186,000 miles a second, while the sound of thunder or that of the whistle comes creeping along at snail's pace, comparatively.

The speed at which sound waves travel depends on the character of the medium through which they pass. Except, perhaps, in the case of violent explosions[1] all sounds travel at equal speed in a given medium. Unless this statement were true, the widely different sounds of the various instruments of a brass band would reach the listener, several hundred yards away, in different sequence from that in which they occurred, with disastrous effect from the musical standpoint.

More specifically, the velocity of sound depends on two characteristics of the medium—its density and its elasticity. The velocity increases directly as the square root of the elasticity and decreases directly as the square root of the density. In the atmosphere, changes in barometric pressure produce no effect on the velocity at which sound travels, so long as the temperature remains the same, the reason being that an increase of pressure increases both the density and the elasticity of the air in the same proportion and *vice versa.*

Hence it is that on mountain tops and at sea level the velocity of sound remains the same if only the temperature does not

[1] Ganot, "Physics," translated by Atkinson, 17th ed., p. 218.
Barton, "Textbook of Sound," pp. 513–553, London, 1908.

change. An increase in temperature of free air, however, decreases the density of the air without affecting its elasticity, and thus an increase of temperature is accompanied by an increase of the velocity. Hence in order to determine velocity of sound in the atmosphere we are not concerned with barometric pressure or with elevation but only with the temperature.

Sound travels faster in solids than in air, but this is not because of the greater density of the solid, as might be assumed. That in itself would reduce the velocity. The greater velocity in solids results from the fact that in these bodies the increase of elasticity over that of air is relatively greater than the increase in density.

At just freezing temperature, 32° F., sound has a velocity in free air of 1,090 feet per second and at 70° F. about 1,132 feet. Starting with 1,090 feet at freezing point, it is a fairly close approximation to say that the velocity increases about 1.1 feet per second for each Fahrenheit degree of rise, or nearly two feet per second for each Centigrade degree of rise.

For the purposes of telephony we are mainly concerned with air as the medium through which sound reaches us. We are constantly submerged in air and cannot live without it. Any sounds which reach our ears through other media are of distinctly minor importance.

But the sounds which come to us through the air are of almost infinite variety, ranging from faint to loud, low to shrill, harsh to sweet, from the rumble of distant thunder to the chirp of a cricket, from the simple hum of a tuning fork to the complex crash of an orchestra and from those which give pain to those which give pleasure. How may we characterize this apparent chaos of differences? Obviously, the speed at which the sounds travel has nothing to do with it, since it has just been shown that nearly all sounds pass through a given medium at the same velocity.

The answer, on the surface, is surprisingly simple. Sounds differ from each other in three characteristics only: *pitch, loudness* and *quality*, or *timbre*. Each of these, in turn, depends on comparatively simple conditions, or combinations of conditions, the principles of which are not difficult to understand.

Simple Harmonic Motion.—As a preliminary to the discussion of these characteristics of sound we may consider briefly the subject of simple harmonic motion. This may appear at first to

be a digression, but in reality it is the most direct approach to the subject.

Simple harmonic motion lies at the root of the whole subject of sound vibration. It is the simplest form of periodic vibration, and it is probably not too broad a statement to say that it is the basis of all sounding vibrations, whether those of noise, speech or music. Any sounding body producing a *tone*, whether a tuning fork, the string of a violin, the reed of a clarinet, the human vocal organs or the air column of an organ pipe, vibrates either with a simple harmonic motion or with a combination of various simple harmonic motions. While it is not so easy to prove, this is probably equally true, in the last analysis, in the case of *noise*, whether the sound of an ax, the crash of breaking glass or the explosion of a mine.

The motion of a simple pendulum vibrating in a small arc is an example of simple harmonic motion. While the pendulum is not ordinarily a sounding body, it may well be considered in gaining fundamental ideas, because its amplitude of vibration is great enough and its rate of vibration slow enough to enable its movement to be observed by the eye, which is usually not the case with sounding bodies.

The pendulum bob, having been deflected from its normal position and freed, starts slowly by gravity to return, gathering velocity until the normal position is reached. At this point its velocity is at its maximum but gravity can no longer urge it forward. It has, however, accumulated kinetic energy, which can be dissipated only by a continued movement against the force of gravity. The movement, therefore, continues with decreasing velocity until all the kinetic energy is converted into potential energy, when it comes to rest beyond the normal point and at the same distance from it as that of the point from which it started. Here, under the influence of gravity, the pendulum starts back in the other direction, and the process is repeated indefinitely. It is characteristic of this motion that it is vibratory and periodic, that it is symmetrical with respect to an axis, that the velocity is always a maximum when passing through the median point, which is the point of equilibrium, and is a minimum, or zero, at the points of maximum deflection.

The forces (Fig. 28) acting on the pendulum bob to return it to its normal position are the pull of gravity acting downwardly and the pull of the supporting string acting toward the point

of support. The resultant of the two acts tangentially to the arc of travel. The horizontal component of this tangential resultant will be equal to the whole force to a very close approximation as long as the deflection is small. Further, the displacement is proportional to the sine of the angle of deflection. The restoring force is proportional to the sine of the angle of deflection, therefore, and, in turn, to the amount of deflection. Hence we have a case of force proportional to displacement. This is the law of force variation necessary to produce simple harmonic motion. In any body vibrating with simple harmonic motion the force produced by its displacement from its normal position is always directly proportional to the amount of its displacement.

The force acting to return the body need not be gravity. More often it is elasticity, as when a stretched string or a reed is plucked, or when air is thrown into vibration by some sounding body. Within ordinary limits the force due to elasticity varies in the manner just described. When a body is displaced from its normal position against elasticity, the force tending to return it is directly proportional to the amount of displacement. As the body moves toward the median point it gathers velocity, thus gradually converting its potential energy, due to its displacement against the elastic force, into kinetic energy, due to its motion. At the median point the velocity is a maximum and all the potential energy has been converted into kinetic. This carries it beyond the normal position, the motion continuing until the kinetic energy is all gone and again replaced by potential energy. Here it stops and starts back again, to repeat the cycle indefinitely.

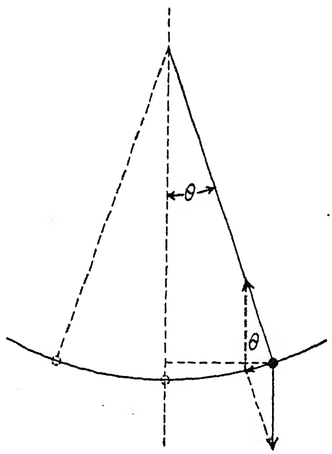


FIG. 28.—Forces acting on pendulum.

The law of variation of the velocity of a point engaged in simple harmonic motion also deserves consideration. It is not sufficient merely to say that it starts slowly from a point of rest, gradually speeds up till it reaches the middle point and then comes to rest again at the point of maximum displacement

on the other side. The variations in velocity occur according to a very definite law. In order to make clear this, and other properties of this form of motion, another illustration will be given:

Consider a pinhead (Fig. 29) projecting from the face of a circular disc and revolving with it at uniform speed on a fixed horizontal axis. The motion of this pinhead will, of course, be uniform rotary motion in a vertical plane. If now a light be placed a great distance away on the level of the axis of rotation and in the plane of rotation, then the motion of the shadow of the pinhead on a vertical plane beyond will be simple harmonic motion. The path of the shadow will be a straight line, the length of which is that of the diameter of the circle in which the pinhead travels. The shadow will move to and fro along this line, coming to rest at each end and moving at maximum velocity as it is passing the center.

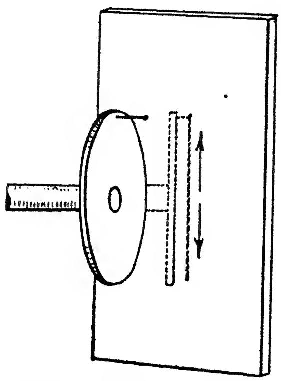


Fig. 29.—Shadow of pinhead moving in circular path.

This particular illustration of the moving shadow does not, of course, define simple harmonic motion. It is merely another example of it, but a particularly useful example, because it points clearly to the close relationship between the properties of the circle and those of this form of vibratory motion. On the basis of it we may go a step further toward definition and say that the movement of the projection on a fixed straight line of a point that is moving uniformly in a circular path is simple harmonic motion.

Let us take another example—that of a "circular pendulum" (Fig. 30) in which the bob travels in a circular path. Here we have a case of uniform circular motion in a horizontal plane. If one views the movements of the bob from a point directly above its point of support, its path will appear as a circle; if the viewpoint is from the side and above the horizontal plane of motion, it will appear as an ellipse; but if from the side and in the plane of motion, the movement will appear to be in a straight line. In this latter case the observed image of the bob will be vibrating to and fro with simple harmonic motion. This brings out, in a simple way, the close relationship between uniform motion in a circular path and simple harmonic vibration

in a straight-line path. One has only to change the point of observation to accomplish the conversion.

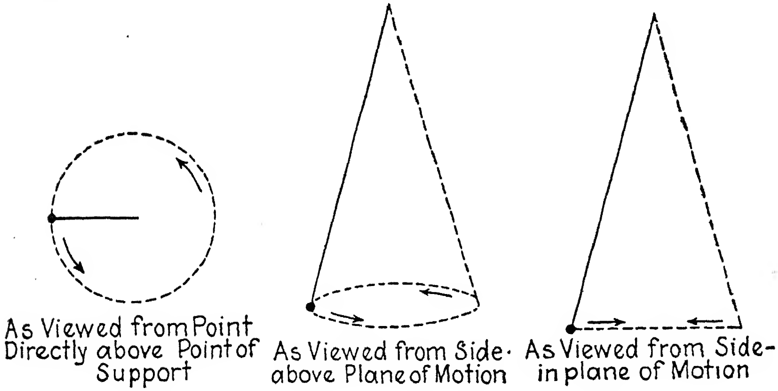


FIG. 30.—Path of circular pendulum bob as viewed from different points.

It is convenient in studying vibratory motion to allow the vibrating object to draw its own path, so to speak, as, for instance, would be done if a pendulum bob (Fig. 31) carried a light pen which traced a mark on a horizontal sheet of paper that was being drawn with uniform motion at right angles to the plane of vibration of the pendulum.

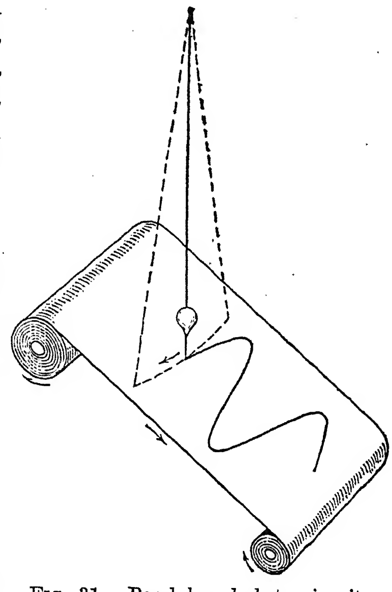


FIG. 31.—Pendulum bob tracing its path on moving tape.

Such a curve, representing harmonic wave motion, may be constructed point by point as shown in Fig. 32. Here a point is considered as moving with uniform velocity around the "circle of reference" in a counterclockwise direction. The projection of this point on the vertical line at the right therefore vibrates along this line with simple harmonic motion. If we divide the circle into any number of equal parts (16 in this case) and measure off, to any convenient scale, a like number of equal lengths on a horizontal

axis, we may consider each such horizontal space as representing the time required for the point on the circle to move through one division on the circle, and for the projected point to move through its corresponding distance on the vertical line. By coördinating each of the successive vertical distances of the point, above or below the axis, with its corresponding time position measured along the axis, the curve at the right of Fig. 32 is traced.

This curve shows the displacement of the vibrating point from its normal position at any instant. It coordinates the movements of the vibrating body with the passage of time. Thus, after the lapse of time represented by, say, two divisions on the horizontal axis, the vibrating point will have been displaced by the distance represented by the ordinate 2-2.

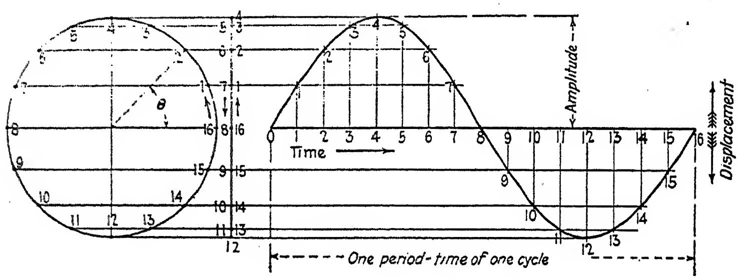


Fig. 32.—Development of sine wave.

Figure 32 may be used as the basis for several definitions important in sound: The *amplitude* of vibration in simple harmonic motion is the maximum amount of displacement of the point from its position of rest. It is half the extreme range of vibration. On the curve it is represented by either of the ordinates 4-4 or 12-12. On the reference circle it is obviously the radius.

The return of the rotating point to the position from which it started completes the *cycle*, which is repeated indefinitely as the rotation continues. With respect to the vibrating point, the cycle includes two returns to the starting point, one while going in one direction, the other while going in the opposite direction. A *cycle* then, may be defined as one complete round of events—*i.e.*, the movements from the time when the body passes a point going in one direction until it passes it again in *the same* direction. It thus includes two single vibrations, one in each direction from the median point. With respect to

the curve traced by the point, the cycle includes one crest and one trough.

The *period* is the time required for one cycle. It is the period of time which the vibrating body takes between two successive passings of the same point in the same direction. In the diagram of Fig. 32 it is the time represented by the distance from 0 to 16 on the horizontal axis.

The number of cycles occurring in one second is called the *frequency*. This is the number of complete to-and-fro vibrations occurring in one second, and is one of the principal factors determining the characteristics of sound. Obviously, it is the reciprocal of the period.

The term *phase* is applied to the portion of a period that has elapsed since a given instant in that period. The initial instant of reference is often taken as that when the vibrating point last passed through the middle of its path in a positive direction. Thus, referring to Fig. 32, we may say that at the instant represented by the point 4 on the curve the vibration has progressed a quarter of a cycle. Or we may say that the point 4 is a quarter phase later than the point 0; or that these two points are a quarter phase apart.

Again, the phase may be expressed in angular measure with respect to the angle through which the point moving on the circle of reference has passed at the instant in question. Thus in passing from 0 to 4 the point would have passed through a phase of 90 degrees; or, as another example, the points 2 and 14 may be said to have a phase difference of 270 degrees.

Such a curve as that of Fig. 32 is called a "sine wave." It is apparent that its ordinates, such as 2-2, 4-4, etc., will always be proportional to the sine of the angle θ through which the vector to the point on the reference circle has revolved at any corresponding instant.

In the analysis of harmonic vibrations it is not always necessary to construct the curve showing the moment-to-moment variations in the position of the vibrating point. Once the conception of the corresponding uniform rotary motion of a point on a reference circle is gained, the entire phenomenon may be represented merely by such a vector diagram as that of Fig. 33, without actually drawing the sinusoidal curve.

In this figure the circle of reference is described on the diameter AB which represents the path along which a mass (or other

quantity) represented by the point P is vibrating about its neutral position O. The force tending to restore the point to this position of equilibrium is always proportional to the displacement of the point from that position. Under these conditions the point will vibrate with simple harmonic motion and its horizontal projection p on the circle of reference will move around that circle with uniform velocity.

Now on this diagram, which is the simplest example of a vector diagram, the following facts are either obvious because of the conditions assumed, or they may easily be proved mathematically:

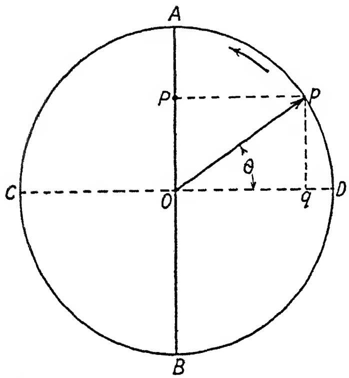


FIG. 33.—Vector diagram of simple harmonic motion.

The force acting to restore the point is always proportional to the distance OP and, therefore, to the sine of the angle θ. It reaches its maximum value at points A and B and is zero at the median point O.

The velocity of the point is always proportional to the distance of the projected point p from the path AB, that is, it is proportional to the line Pp and, therefore, to the cosine of the angle θ. It is zero at points A and B and a maximum at point O.

The acceleration of the point, which is the rate at which its velocity is changing, varies in exactly the same manner as the force, and, hence, is proportional to it, as in all mechanical phenomena $\left(\text{acceleration} = \dfrac{\text{force}}{\text{mass}}\right)$.

The period of oscillation and also its reciprocal, the frequency, are independent of the amplitude of motion: that is, independent of the length OA or OB.

The energy of oscillation is represented by the kinetic energy stored in the mass when passing its position of equilibrium and is proportional to the square of the amplitude OA and, at the same time, to the square of the frequency.

The fact that simple harmonic motion in a straight line may be referred so easily to the motion of a point traveling at uniform speed in a circular path is of deep mathematical significance. It affords at once a convenient basis for either the mathematical

or the graphical analysis not only of all harmonic motion but also of all periodic phenomena whose variations occur in accordance with the curve of sines. It applies, therefore, to *all sound waves* and also to all *alternating current waves.*

As has already been pointed out, telephony involves the conversion of sound waves into alternating current waves, the transmission of these current waves to a distant point, and their reconversion at that point into sound waves. The laws of simple harmonic variation therefore apply to the electrical quite as much as to the acoustic phase of telephony.

The fact that we can reason about simple harmonic motion or about harmonic variations, such as those of current or electromotive force, in terms of uniform motion in a circular path accounts for the frequent occurrence in mathematical treatises on sound and on alternating currents of such symbols as $2\pi r$ and of sine, cosine and other trigonometric functions. The subject of harmonic vibrations can be most effectively dealt with in a mathematical way. This is not a mathematical treatise, but those who feel a desire to delve into the mathematics of the subject will find a large field of opportunity and a wealth of literature treating of it.

Complex Periodic Vibration.—So far we have considered the simplest manner in which a body may vibrate. Such simple vibrations, however, are rarely found alone in nature. As a rule the vibrations are far more complex for the reason that bodies usually vibrate not only as a whole but also in various parts. For this or other causes smaller ripples of different sizes are usually superposed on the main wave. The resulting wave form is correspondingly altered.

Fourier's theorem tells us, in effect, that any periodic curve, no matter how complex, may be built up by compounding a number of simple harmonic curves properly chosen with respect to their amplitudes, periods and phases. Conversely, any complex periodic curve may be resolved into a definite number of simple harmonic curves of definite periods, amplitudes and phases. It is also a part of Fourier's theorem that the periods of all the simple harmonic curves must be exactly commensurate—that is, they must all be exact divisors of the fundamental period. This theorem is true not only of the curves themselves but also of the motion or any other phenomenon which the curves may represent.

Since the period is always the reciprocal of the frequency, it follows that every complex periodic vibration may be definitely

resolved into a number of simple harmonic vibrations, the least rapid of which has what we call the "fundamental frequency" and all of the others have frequencies that are exact multiples thereof, no fractional multiples existing.

We may gain an exact conception of the nature of such complex periodic vibration by considering the diagram of Fig. 34. Here, as in Fig. 33, the point p revolving uniformly in the large circle is projected, as by a shadow, onto the vertical diameter of that circle. The projected point P then oscillates along this diameter with simple harmonic motion. This motion, since it has the lowest frequency of the series to be considered, constitutes the *fundamental* wave motion of the series.

The point p on the large circle forms the center of another circle around which the point q moves with uniform velocity with just half the period and twice the frequency of that of the first point, p. Considered alone, the projection of this point q will move up and down the vertical diameter of its own circle with simple harmonic motion. This constitutes the second wave motion of the series and, since it has a frequency twice that of the fundamental wave, it is called the *first harmonic* of the fundamental wave.

The whole circle in which q revolves is at the same time moving bodily around the large circle which forms the path of p. While the projection of q is moving up and down, therefore, due to its motion in its own orbit, it is also partaking of the up and down motion of p in its larger orbit. Obviously, the motion of Q, the projection of q on the vertical line AB, will be a combination of the two harmonic motions, the fundamental and its first harmonic.

To go a step further, the point r revolves in an orbit of its own about the center q with a period one-third and a frequency three times that of the point p. The projection of r on the vertical diameter of its own orbit will thus constitute the third wave motion of the series. It has just three times the frequency of the fundamental and is called the *second harmonic* of the fundamental wave.

As the center of the orbit of r is also partaking of the movements of the other two points p and q, it is clear that its projection R on the vertical line AB will move with a complex periodic wave motion which is the resultant of three component simple harmonic motions whose respective frequencies are as one, two

and three, and whose amplitudes are the radii of the respective circles in which the three points move.

In Fig. 34 the three revolving points are shown as starting in the same phase; in the full-line circles each is on the common horizontal axis. Let us consider what takes place as the movements progress. At the instant when the fundamental wave has advanced 30 degrees in phase, or one-twelfth of a cycle, the point p has moved to p' on the large circle, and its projected point has had a corresponding movement from P at the center of the large circle to P' above it.

During this time the point q, rotating with twice the angular velocity, has advanced 60 degrees on its circle, and also the whole circle in which q revolves will have moved up to the position shown in dotted line. The point q will, therefore, have moved to q' on its dotted circle, and its projected point on the vertical axis will have moved from Q to Q'.

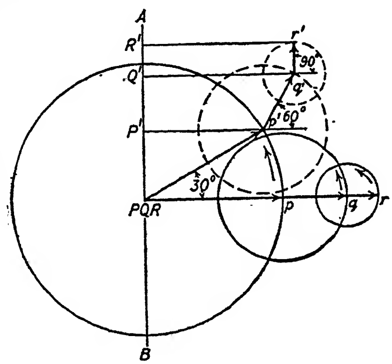


FIG. 34.—Combination of three simple harmonic motions.

At the same time point r moving with three times the angular velocity of p will have advanced 90 degrees on its circle, that circle also having moved into a new position as shown by dotted line. The new position of the point r is, therefore, at r', and its projected point has moved from R to R'.

It is to be noted that the diameter of the circle of reference of the fundamental movement requires extension, since the displacements in the complex wave will often exceed the amplitude of the fundamental wave. The resulting displacement (from O) of the point R is always the projection of the vector sum of the three radii carrying the respective points whose motions are contributing to the complex wave.

If the three circles of Fig. 34 are considered as cog wheels geared together on the centers and with the respective ratios mentioned, then the shadow R which a pin at r would cast on the vertical line AB would move with the complex periodic motion produced by the three simple wave motions. This conception may be extended indefinitely by considering as

many additional orbits or cog wheels as may be required to represent all the components or higher harmonics constituting the complex periodic wave form under consideration. The radii of the added wheels will correspond respectively to the amplitudes of the components, and their speeds, always geared to exact multiples of that of the slowest wheel, will represent respectively the successively higher frequencies of the series of harmonics comprised in the complex wave.

We may graphically coordinate with time the displacements of these three harmonic motions and their resultant as in Fig. 35. This is drawn to the same scale as Fig. 34 to enable ready

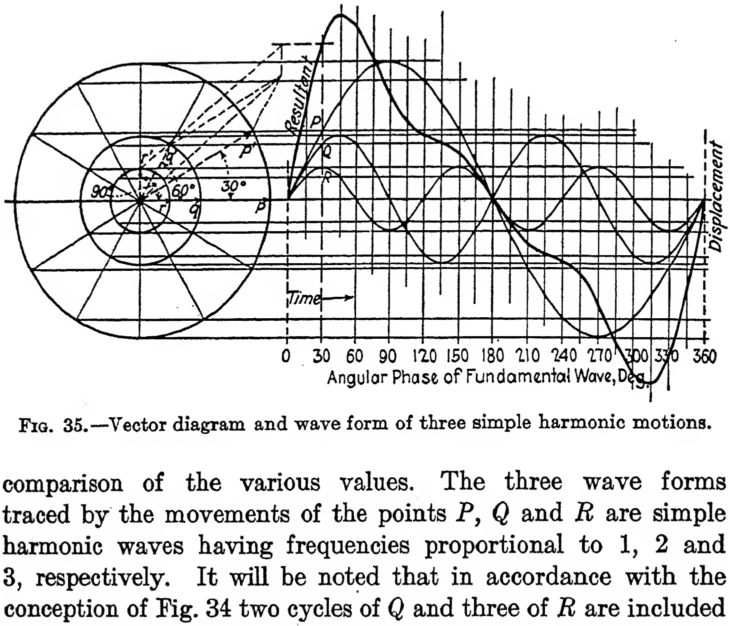


FIG. 35.—Vector diagram and wave form of three simple harmonic motions.

comparison of the various values. The three wave forms traced by the movements of the points P, Q and R are simple harmonic waves having frequencies proportional to 1, 2 and 3, respectively. It will be noted that in accordance with the conception of Fig. 34 two cycles of Q and three of R are included in a single cycle of P. The heavy resultant wave form is found by taking for each point on it the algebraic sum of the corresponding ordinates of the three component waves.

An examination may be made of the conditions at any phase of the cycle, either by studying the wave forms at the right or the vector diagram at the left of Fig. 35. The conditions indicated by the arrows on the vector diagram are those at the instant when the fundamental wave has advanced 30 degrees in its cycle. The value of the resultant displacement at this instant may be found by treating the arrows as though they were forces, combining as by a parallelogram of forces the values of

p' and q' to obtain the displacement at that moment due to the fundamental and first harmonic waves P and Q, respectively,

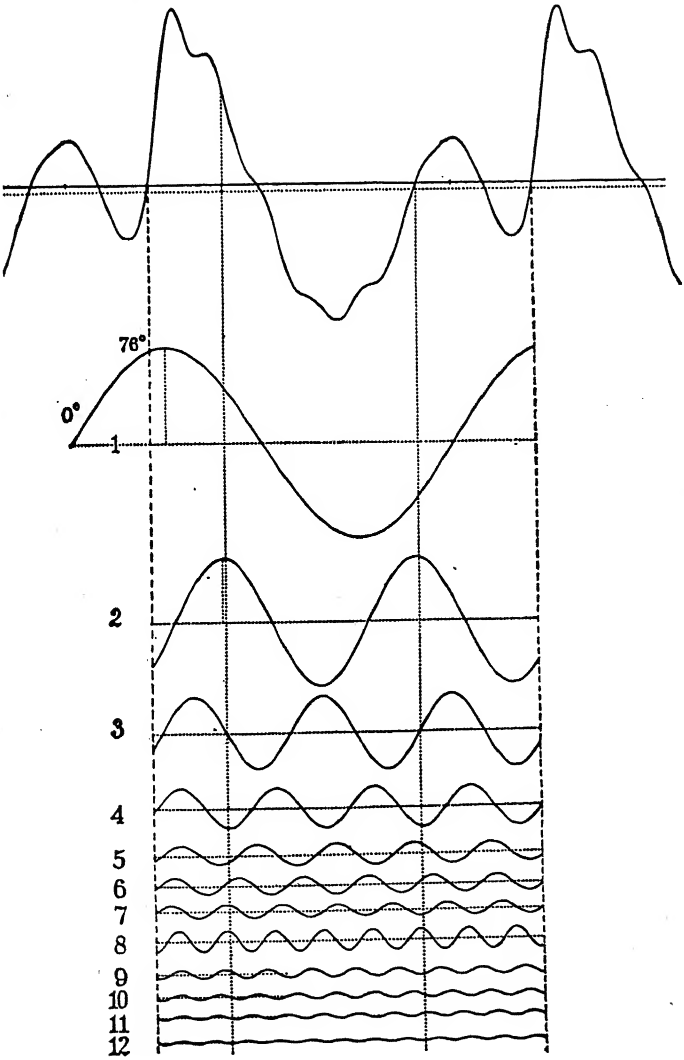


Fig. 36.—An organ-pipe curve and its harmonic components.[1]

and then forming a parallelogram of this resultant with the arrow r' to obtain the instantaneous displacement due to the

[1] Reprinted from Miller, D. C. "The Science of Musical Sounds," The MacMillan Company.

combined action of all three waves. The same result is found by adding the ordinates of the three component waves on the 30-degree phase.

As a somewhat more extended analysis of wave form, Professor Dayton C. Miller has shown the fairly complex form at

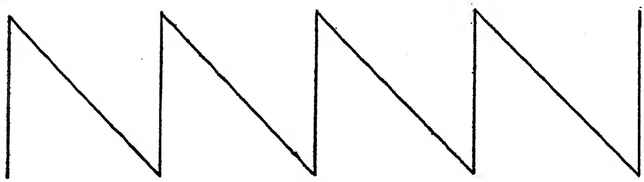


FIG. 37.—Saw-tooth wave.[1]

the top of Fig. 36 to be composed of the twelve simple harmonic components shown below it.

In view of this, one may ask concerning the limitations of this ability to produce complex periodic wave motions by combining simple harmonic motions. Within finite limits there are none. Any periodic wave form, whether it consists of

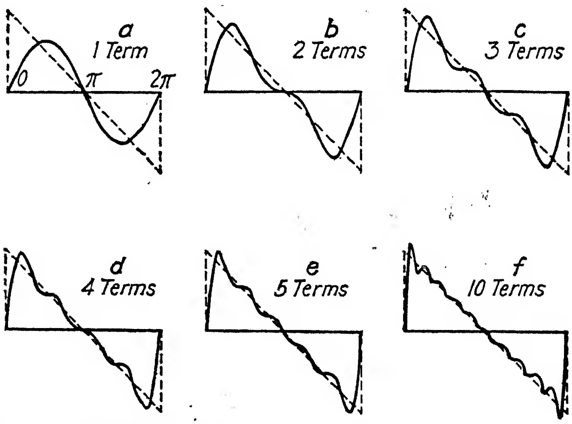


FIG. 38.—Approximations to saw-tooth wave obtained by compounding harmonics up to tenth.[1]

combinations of curved or of straight lines, may be so compounded if only enough properly chosen simple harmonic waves are employed. Figures 37 to 39, also taken from Professor Miller's work, show a striking example, that of building up a sawtooth wave form, consisting wholly of straight lines and sharp angles, by combining simple harmonic curves. Figure

[1] Reprinted from MILLER, D. C. "The Science of Musical Sounds," The MacMillan Company.

37 shows the sawtooth wave form to be synthetically compounded. Figure 38 shows successive steps leading to closer and closer approximations as additional sinusoidal wave forms are added. Figure 39 is the approximation arrived at after 30 terms of the series[1] had been added. An infinite number of terms would exactly reproduce the figure.

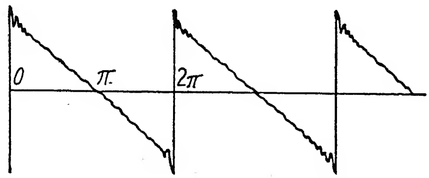


FIG. 39.—Approximation to saw-tooth wave obtained by combining 30 harmonics.[2]

Graphic Representation of Sound Waves.—We may employ such curves as those of Figs. 32 and 35 in the analysis of sound waves with two different conceptions in mind. In one, which was that used in the discussion of Fig. 32, the successive points on the curve represent the successive displacements of the vibrating body as time progressed. In the other, applicable particularly

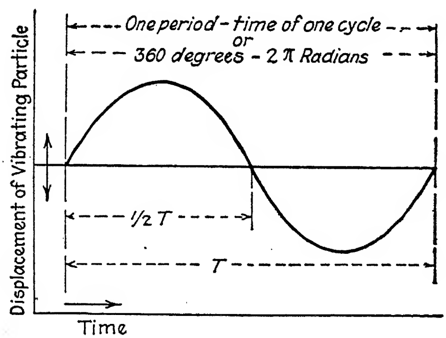


FIG. 40.—Displacements of vibrating particle with lapse of time.

to sound waves in air, the successive points on the curve may be considered as indicating the respective displacements *at a given instant* of the successive air particles engaged in the wave. The curves of Figs. 40 and 41, illustrating these two conceptions, are exactly alike, the notations only differing. The essential differ-

[1] To represent exactly the sawtooth form of Fig. 37 the successive component waves would be represented by the infinite series:

$$y = 2(\sin x + \tfrac{1}{2}\sin 2x + \tfrac{1}{3}\sin 3x + \tfrac{1}{4}\sin 4x + \cdots).$$

[2] Reprinted from MILLER, D. C. "The Science of Musical Sounds," The MacMillan Company.

ence in the two conceptions is that, in one, horizontal distances represent the passage of time, in the other, the traversing of distance. Either the time or the distance may be measured in any appropriate units. For instance, in the conception of Fig. 40 time may be expressed in seconds, or in any function directly dependent on time, such as the passage of the imagined point on the circle of reference expressed in angular measure or in radians measured around the circle of reference.

In Fig. 41, as before, the ordinates above or below the horizontal axis represent displacements. The successive points along the curve represent, therefore, the positions of the successive particles engaged in the wave all at the same instant of time.

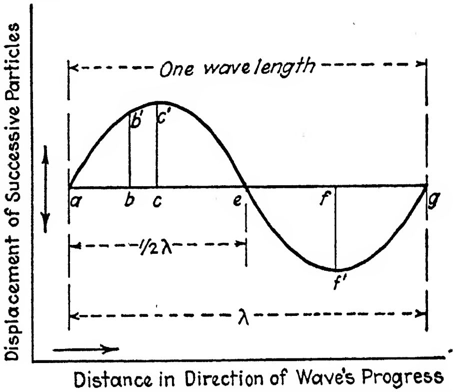


FIG. 41.—Positions of successive particles in wave at any single instant.

Here, then, we have the representation of a complete simple harmonic wave in air. Starting at *a* the particle of air would be in its neutral or normal position. Passing further along, a distance *ab* from the source of the sound, the displacement of the particle normally at *b* would be indicated by the ordinate *bb′*. At the distance *ac* the displacement would be represented by *cc′* and so on; the particles at *a*, *e* and *g* having no displacement, those at *c* and *f* having maximum displacement.

This diagram of Fig. 41 we may consider as representing the cross-section of a wave at a given instant. Obviously, then, the distance *a-g* is the *wave length*—that is, it is the length between corresponding points on successive waves. Looking at it in another way, it is the distance that the wave front travels for each complete to-and-fro vibration of the sounding body, or for each cycle.

Obviously, there is an exact relationship between wave length, frequency and velocity in a given medium. If the frequency is known as so many cycles per second, then the sound must travel that many wave lengths in a second; *i.e.*, the product of wave length and frequency must be the speed at which sound travels. The wave length, therefore, in a given medium, is the speed of sound in that medium divided by the frequency. The Greek letter λ (pronounced lamda) is commonly employed as the symbol for wave length. The relationship between wave length, velocity and frequency is thus expressed algebraically:

$$v = f\lambda \text{ or } \lambda = \frac{v}{f}.$$

To illustrate by a few examples of familiar sounds occurring in air: The note of a very low organ pipe may have a frequency of 20 cycles per second. We know the speed of the sound to be about 1,132 feet per second. Therefore, the wave length of that sound will be 56.6 feet. Going to the other extreme and considering a sound so shrill as to be scarcely audible, the rate of vibration or frequency would be perhaps 10,000 cycles per second, in which case the wave length would be only about 1.36 inches. Middle *c* on the piano, with a frequency of 256 complete vibrations a second, would have a wave length of 4.4 feet.

Contrast these wave lengths with the amplitude of the waves. These are nearly always measurable in very small fractions of an inch—in fact it is known that vibrations less than one-millionth of an inch in amplitude produce sounds readily discernable by the human ear.

Whether we consider the curve as representing the successive displacements of the *same* particle of air with the passage of time, as in Fig. 40, or as representing the displacement of *successive* particles in a wave at a given instant, as in Fig. 41, it must be kept in mind that the vibrations in a sound wave are longitudinal, and the displacements of the particles in such curves would all be either to the right or to the left along the horizontal axis. In order to represent them graphically on a surface, however, the distances through which the particles have moved are shown at right angles to the directions in which the movements actually occur, those displacements in a positive or forward direction being shown above the axis, those in a negative or backward direction below the axis.

In order to secure visible magnitudes in our diagrams it is necessary to plot the vertical distances representing displacement on a greatly enlarged scale. The actual displacements caused by a sounding body are usually very minute, often of the order of one hundred-thousandth of an inch, whereas the actual wave length, as has just been shown, may vary from an inch or less up to perhaps a hundred feet. If we attempted to lay out such a curve to actual scale, it would not differ visibly from a straight line.

Simple and Complex Sound Waves.—We have seen from Fourier's theorem that every periodic wave form consists either of a single simple harmonic wave or of a number of simple harmonic waves, of which one, the fundamental, has the lowest frequency, and the others, the harmonics, have frequencies respectively two, three, four, five . . . etc., times as great, no fractional multiples being found. In acoustics a simple sound wave consisting merely of a single simple harmonic oscillation produces a sound which we call a *simple tone* or a *pure tone*. On the other hand, a complex sound wave, consisting of a number of simple harmonic component waves, produces a *complex tone*.

Fourier's theorem as applied specifically to acoustics brings us, therefore, directly to the statement that every complex tone consists of a fundamental simple tone and its harmonics, which are also simple tones.

Tones, whether simple or complex, always result from *periodic* vibrations. *Aperiodic* vibrations, that is, those whose wave forms do not apparently recur at regular intervals, produce what we call "noise." This distinction between tone and noise is not altogether complete or satisfactory and will be dealt with more fully in subsequent chapters.

Probably the tone of a tuning fork comes as near to representing a single train of simple harmonic vibrations as any familiar sound. It produces a rather sweet but uninteresting insipid tone. It is lacking in the harmonics or "overtones" which add richness or character to the tone produced by the fundamental when sounding alone. On the other hand, the violin produces excellent examples of complex tones.

We may gain an idea of what happens in the case of more complex vibration by considering a tightly stretched string, such as the *G* string on the violin. If the bow is applied to the open string, a note characteristic of that string is heard. By

watching the string it may be seen that it is vibrating as a whole from one end to the other, in a manner indicated at *a* (Fig. 42). As a matter of fact, the string, while vibrating in this relatively simple way, is also vibrating in other ways, far more complex, that cannot so easily be visualized.

One of these other modes of vibration may be illustrated by applying the tip of the finger very lightly to the center of the string so as to prevent its vibrating at that point and again applying the bow. The string will now vibrate in half lengths, as indicated at *b* (Fig. 42). A *node* will exist at the center point which is *stopped* by the finger. The vibrations, being due to the half length of the same string at the same tension, will have double the frequency of those of the open string. Musicians

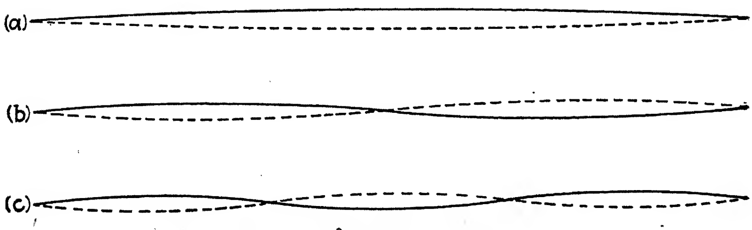


Fig. 42.—Vibrations of string.

will recognize the tone now produced as being the *octave* of the tone of the open string.

In a similar manner a skilled violinist, by touching the string lightly at other points along its length, may make it vibrate in thirds, as at *c* of Fig. 42, or in quarters, fifths or other aliquot parts of its length. Each time a different sort of tone would be produced, always higher than that of the string vibrating as a whole.

The tone emitted by the string when vibrating as a whole, as at *a*, is its fundamental tone, it being the lowest tone of which the string is capable at that particular length and tension. The tones which it emits on account of the independent vibrations of any of the aliquot parts of its length, as at *b* and *c*, are the harmonics or overtones of that string. The production of these harmonic tones on violin strings by lightly touching the string at certain points, as just described, is an important and difficult part of the technique of the violinist. It is mentioned here because it illustrates clearly how the fractional parts of a sounding

body may be made to take up vibrations independent of its vibrations as a whole.

When a violin string is bowed, it automatically takes up these various vibrations corresponding to its whole length and its various fractional lengths, without any attempt on the part of the violinist to make it vibrate other than as a whole. As a result the various harmonics are present with the fundamental as component parts of the whole complex tone. Because the harmonics are parts of the whole tone they are often referred to as *tone partials.*

The motion of any point on the string as the result of all of these different vibrations going on simultaneously is very

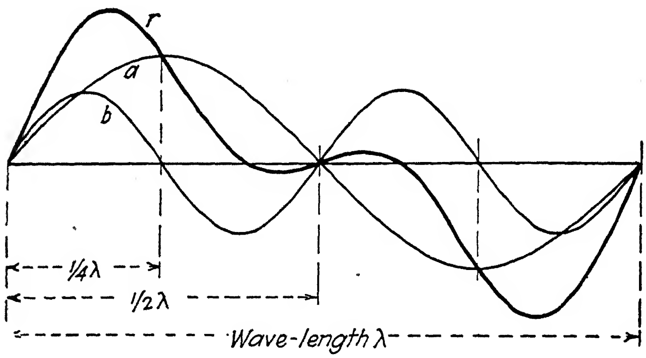


Fig. 43.—Combination of two harmonic waves.

complex. It is subject to ready analysis, however, if only the frequencies, amplitudes and phase relationship of the various component simple harmonic waves are known. With these determined one has only to combine the component waves into a resultant wave in exactly the manner illustrated in connection with Fig. 35.

The resultant wave form of that figure, for instance, might be considered as representing the moment-to-moment displacements of a point on the violin string when the string was vibrating as a whole (movements of point P), and as to its two half lengths (movements of point Q) and as to its three third lengths (movements of point R). It is to be remembered that the resultant displacement of the point from its normal position at any instant, due to the three independent vibrations, is found by taking the algebraic sum of the three displacements at that instant.

Let us now inquire briefly how changes in the component waves of a complex sound may alter the resultant wave form. This is pertinent because, as will be shown later, some changes which affect wave form will produce marked changes in the character of the sound sensation, while others do not.

In Fig. 43 is shown the analysis of a very simple case—a wave form composed of only two simple harmonic motions, represented by sine curves a and b with wave lengths of λ and $\frac{1}{2}\lambda$, respectively. The heavy curve r is that of the resultant complex tone. These two component waves are shown in phase with each other, since each starts its cycle at the same instant.

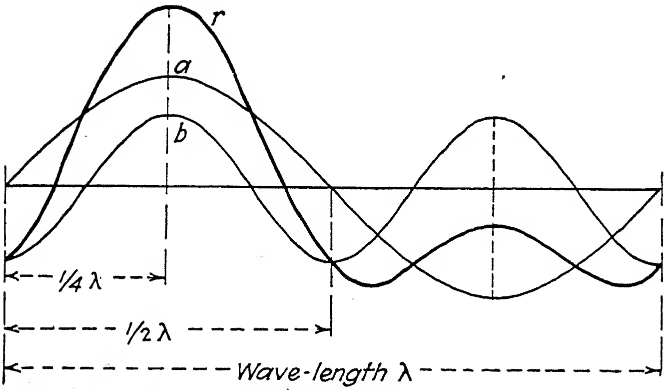


FIG. 44.—Effect on wave form of shifting phase of components.

Figure 44 shows a case where the same two curves a and b, are not in phase, b lagging one-eighth of a cycle or one-eighth of a wave length behind a. Here the resultant wave r, while of the same wave length, has a decidedly different form and also a different amplitude. We see then that the phase relation of the components of a complex sound has an effect on the wave form.

In Fig. 45 components a and b are present, each with the same period and frequency as in the two preceding figures but with different relative amplitudes. In this case the amplitude of a is half of what it was in Fig. 43, while that of b is double. Here, although the phase relation, periods and frequencies are the same as before, the resultant curve r has been changed radically in form. Evidently, therefore, the form of a complex wave may be varied by changes in the amplitude of its components, all other characteristics of the components remaining the same.

As another illustration, in Fig. 46 components *a* and *b* are exactly the same in frequency, phase and amplitude as in Fig. 43, but another simple harmonic component *c* has been added. This new component has smaller amplitude and four times

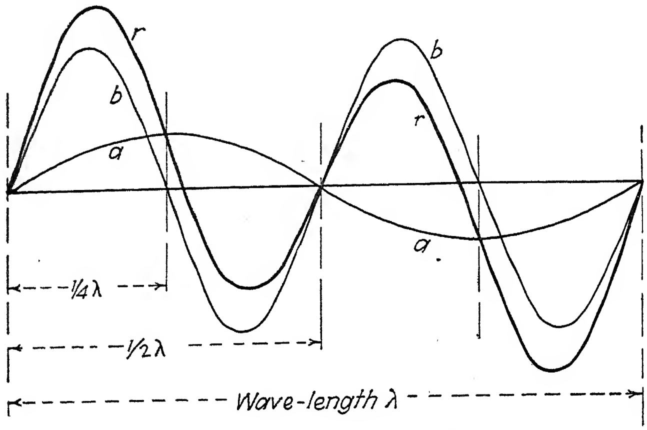


FIG. 45.—Effect on wave form of varying amplitude of components.

the frequency of the fundamental *a*. Again we see that the resultant wave form *r* has been radically altered by the addition of another component.

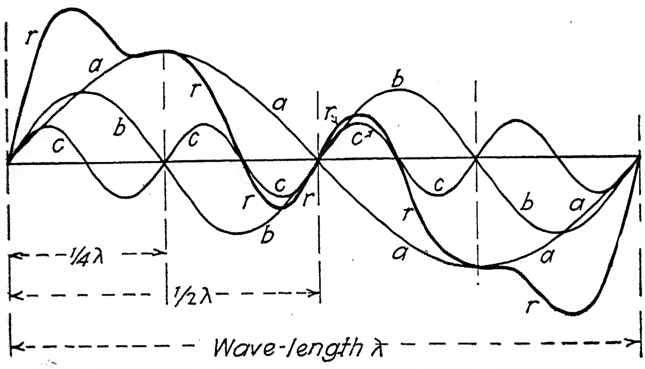


FIG. 46.—Effect on wave form of adding another component.

It is clear, then, that variations in wave form of a complex sound wave may be caused by the relative shifting of phase with respect to the components, by the variation of the amplitude of the components, and by variations in the number of components. As will be shown later, we are not very much concerned

with those changes in wave form which result from a mere shifting of the phase relationship of the components. We are, however, vitally concerned with those differences which are the result of changes in the number and amplitudes of the component simple harmonic waves.

The actual building up of a wave form graphically from its known simple harmonic components is a sort of synthetic process. Knowing the components with respect to their frequency, amplitudes and phase relationships, it is easy to determine

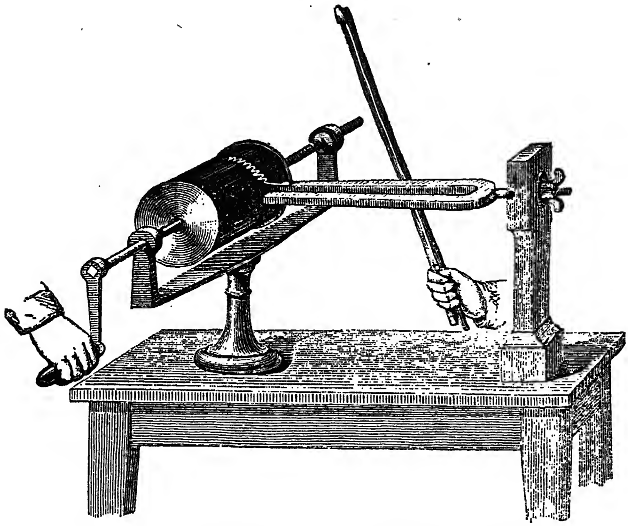


FIG. 47.—Tracing wave of tuning fork vibration.

the form of the resultant wave. The reverse of this process, however, is not so easy. If the graphical form of a complex wave is known, its analysis may be done by a system of mathematical analysis devised by Fourier. Also there are various wave analyzers by which any graphical periodic wave form may be resolved into its simplest components with comparative celerity.

Obviously, then, one way of analyzing an actual sound is to make it draw its own curve and then analyze the curve. A rudimentary way of doing the curve drawing is that suggested in Fig. 31, and shown applied to the vibrations of a tuning fork in Fig. 47.[1] But such direct methods of curve drawing

[1] Reproduced from Deschanel's "Natural Philosophy."

are applicable only to very simple vibrations of comparatively large amplitudes. The more delicate vibrations, particularly of the overtones, are lost in working with so gross a method. Fortunately, methods of far greater refinement are at hand.

There are various "oscillographic" devices for recording sound vibrations. In one of these, shown in principle in Fig. 48, a beam or pencil of light shines through a slit on to a small mirror of minute weight. This mirror is made to vibrate in accordance

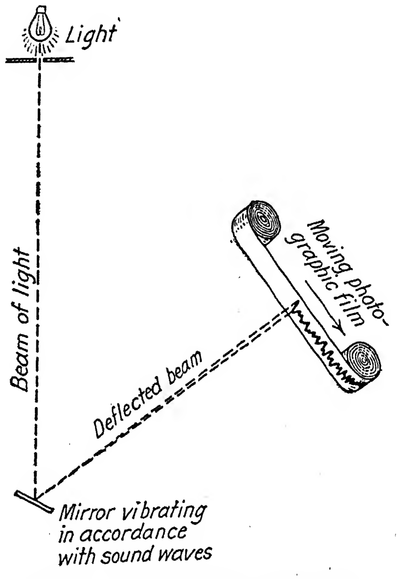


FIG. 48.—Principle of the oscillograph.

with the sound waves to be studied, and in its vibration it causes the reflected beam of light to vibrate sidewise. This reflected beam, shining on a photographic film which is drawn along at a uniform rate in a plane to which the beam is perpendicular, leaves a record on the film. When the beam is standing still in its normal position, it draws a straight line, but when it is thrown into vibration, it draws a wavy line corresponding in its displacements from the normal or axial line with the varying amplitudes of the sound vibration. The mirror may be mounted on the vibrating body and partake directly of its motion or it may be separately mounted and moved indirectly by electromagnetic influence under control of the sound. In other forms of oscillograph a beam of cathode rays is deflected by passing through a

magnetic field, the strength of which is varying in accordance with the sound vibrations.

An oscillogram of a complex sound is shown in Fig. 49. The sound in this case was that of the voice in uttering the word "poor." The lower record of this figure is that of a simple wave at a frequency of 500 cycles per second. The records of this figure have the same horizontal scale, so that the lower one may be used as a guide to determine the frequencies of the upper.

Even with the great sensitiveness of these oscillographic wave tracers, many of the upper harmonics of complex sounds, such as

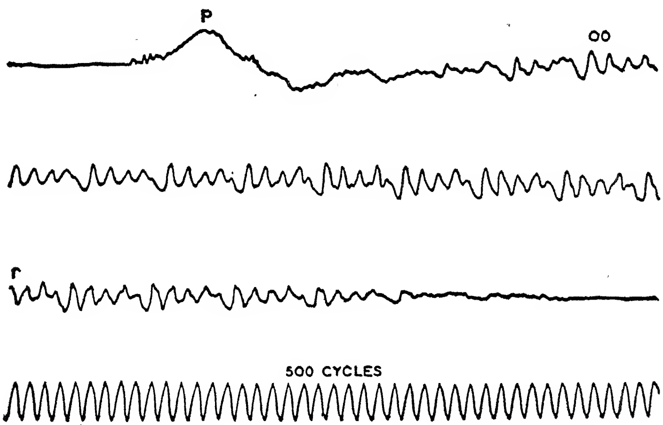


FIG. 49.—Wave form of the word "poor." (*Courtesy of Bell Telephone Laboratories.*)

those of the voice, are lost and, for more complete analysis, methods of still greater refinement are required. These methods employ what are called "harmonic resonators" or "frequency analyzers." They determine directly what frequencies are present in the complex sound instead of first drawing the curve of the sound wave and then analyzing the curve for its component frequencies. In order to understand these direct methods of sound analysis, brief reference must be made to the principles of forced and resonant vibration.

As a general principle, it may be stated that any free vibratory body will have its own natural rate of vibration. Such a body, however, will be set in vibration by *any* periodic force applied to it, whether the frequency of the periodic force coincides with this natural rate or not. When the frequency of the periodic force differs from the natural rate of the body, the vibrations thus

produced are called *forced vibrations.* The amplitude of the vibrations will increase as the frequency of the applied force approaches more nearly to the natural rate of the body and will be a maximum when the frequency of the force exactly coincides with the natural rate of the body. The body will then be in *resonant vibration*, a limiting example of forced vibration. Under conditions of resonance, the increase in amplitude of vibration is most pronounced. Indeed, if there were no dissipation of energy, through friction or otherwise, the resonant vibration would increase indefinitely with the continued application of the periodic force.

We are all familiar with the fact that a piano string will be thrown into vibration by a near-by sound having a frequency of vibration corresponding to the natural frequency of the string. One tuning fork can be set in vibration by another of exactly the same pitch vibrating near by. These are resonance phenomena.

A striking example of a heavy body being kept in pronounced vibration by the application of minute periodic forces, applied at just the right intervals, is found in a long-pendulum clock. The pendulum, weighing perhaps several pounds, is sustained in vibration by the feeble impulses it receives from the clock escapement, because, and only because, the frequency at which these occur is exactly that of the natural frequency of the pendulum. In the case of the tuning forks the communication between the two is through air; the sounding fork causes waves in the air and these waves act, like the well-timed impulses of the clock escapement, to cause a cumulative vibratory effect on the second fork.

Masses of air, partially confined, have a natural rate of vibration. This may be shown by holding a vibrating tuning fork over a tall glass jar into which water is gradually being poured. As the water level rises, the air column in the jar above it becomes shorter, and its natural rate of vibration is changed accordingly. When this natural rate is thus brought into coincidence with that of the tuning fork, the resulting sound will swell forth loudly. As will be shown in a subsequent chapter, this resonance of air masses forms an important phase of the production and modulation of sounds by the human organs of speech. It has also played an important part in the analysis of sound.

Reverting to the direct methods of sound analysis, Helmholtz, in his pioneer investigations of the character of complex sounds,

used a series of air "resonators." Each resonator consisted of a globe of thin brass with a large hole at one side for the reception of the sound and a small hole and nipple at the other side for application to the listener's ear. Each resonator was of such size and form as to give a particular natural rate of vibration to the air within it—it was tuned to a desired frequency. If the complex sound under investigation had that frequency as one of its constituent harmonics, that tone would be greatly augmented in the resonator. By using a series of resonators, tuned respectively to a series of harmonic frequencies, the presence or absence of any one of these frequencies in a complex tone could be determined by listening for it through the corresponding resonator.

Most frequency analyzers up to a recent date have been based on this principle, involving the resonance of a body of air, but

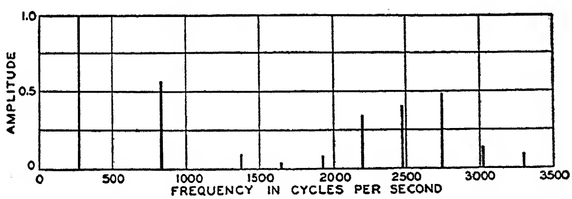


Fig. 50.—Component frequencies of clarinet c′. (*Courtesy of Bell Telephone Laboratories.*)

recently a more accurate method, involving the resonance of tuned electrical circuits, has largely supplanted it.

The principles of these electrical frequency analyzers will be better understood after the subject of tuned electrical circuits has been discussed. For the present purposes, however, it is sufficient to say that they consist in effect of a large number of "electrical filters," each so adjusted as to permit currents of only a single frequency, or of a very narrow band of frequencies, to pass, all other frequencies being almost completely suppressed.

Since these filters are effective only on electric waves and not on sound waves, it becomes necessary first to convert the sound waves into corresponding electric waves and then to analyze these by means of the filters. By a suitable telephone transmitter the complex sound under analysis is made to generate alternating currents which contain all the frequencies of its component simple tones, all with their proper relative amplitudes. By successively applying these filters throughout the entire range of frequencies being studied, the presence of each frequency in a sound may be determined and its relative intensity measured.

The results thus determined, showing the component tones of a given complex sound, may be graphically represented as in Fig. 50. In this the presence of each component of the sound is indicated by one of the vertical lines. The position of each line with respect to the horizontal scale indicates the frequency; and the length, as shown on the vertical scale, the amplitude. Thus we have what may be called a "spectrum" of the complex sound, the one in this case being that of the tone of the clarinet at a pitch corresponding to a frequency of 256 cycles per second.

This method of sound analysis, by determining what frequencies are present and their respective amplitudes, gives directly, and without resorting to curves, the most important information that is required in telephony. It does not give the phase relations of the various components, but, as will be shown, this is unimportant in telephony.

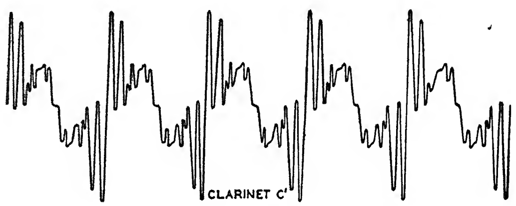


FIG. 51.—Wave form of clarinet c′. (*Courtesy of Bell Telephone Laboratories.*)

If, for any reason, it is desired to draw an actual curve of the wave form, this may be done by combining graphically the simple harmonic curves representing each frequency, as already described, it being necessary only to assume the phase relations. Such a curve drawn from the data contained in Fig. 50 is shown in Fig. 51.

With the understanding, so far attained, of the nature of sound vibrations, we may now study the characteristics of sounds by which our sense of hearing enables us to distinguish them from each other. As stated in the earlier part of this chapter, there are three such characteristics and three only—*pitch, loudness* and *quality*. Quality is often called *timbre* or *tone color*.

Pitch.—Pitch is that characteristic by which we distinguish so-called "high" notes from "low" notes. More particularly, it is the characteristic which determines the position of the tone on the musical scale.

Pitch is solely a function of the frequency of vibration. In a simple tone it is determined by the frequency of the single simple wave. In a complex tone, with exceptions that will be mentioned, pitch is determined by the frequency of the fundamental component wave.

Vibrations of lower frequencies produce the so-called "low" tones; those of high frequency the "high" tones; hence it is that the low notes on a piano are produced by long strings of relatively great mass which vibrate slowly, while the high notes are produced by short strings of small mass so stretched as to vibrate rapidly.

Since frequency alone determines pitch, it is easy to see that the pitch of all the tones indicated in Fig. 52 would be the same

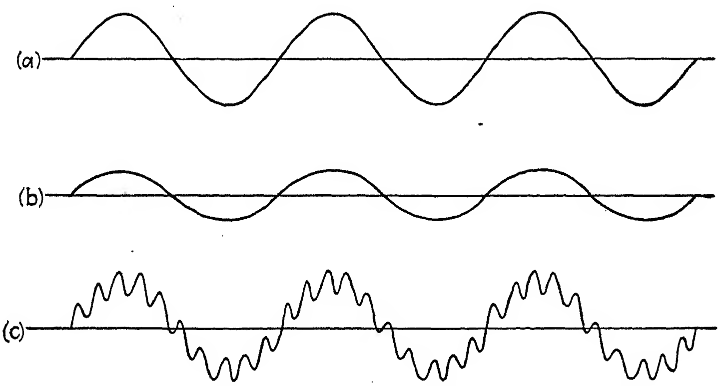


FIG. 52.—Different wave forms of sounds having same pitch.

in spite of their different forms and amplitudes. All three have frequencies of, say, 256 cycles per second. The upper and middle curves represent simple tones with widely different amplitudes. The lower one is of a complex tone having a harmonic of ten times the fundamental frequency compounded with the fundamental. In spite of these differences one's sense of hearing would tell him, without question, that they were all of the same pitch.

It has always been thought that the pitch of a complex tone was determined by the frequency of the fundamental component of the tone. Recent work by Dr. Harvey Fletcher and his associates in the Bell Telephone Laboratories has shown, however, that the fundamental and several of the lower harmonics may

be eliminated from a complex tone without changing its pitch.[1] This is somewhat revolutionary and disturbs the hitherto accepted definition, that the pitch of a complex tone is that of its fundamental, since with the fundamental gone the pitch remains. Evidently, since the sound waves coming to the ear from an external source contained no frequency as low as that of the fundamental of the tone heard, the ear mechanism itself, by some sort of subjective operation, must have introduced a tone corresponding to that frequency. The matter is, therefore, one of interpretation by the ear mechanism and will be treated much more extensively in the next chapter which deals with sound from the standpoint of sensation rather than of physical sound waves *per se.*

For most practical purposes, at least, the old definition, that the pitch of a complex tone is that of its fundamental rate of vibration, may stand. So far as is known it is true of every naturally produced complex tone. It is only when man, either by intent or *by defective design of apparatus or system,* cuts off some of the frequencies that would naturally lie at the base of a harmonic series of vibrations that the ear steps in and supplies the deficiency, with a resulting sense of pitch not in conformity with the generally accepted definition.

The pitch of sounds plays an important part in telephony. For this reason and also because, with the advent of radio, the transmission of music is demanding more and more attention, it is desirable for the telephonist to be able to associate concrete musical tones with their corresponding frequencies and to understand something of the way in which these tones and frequencies are related in the musical scale. It is also well for him to be able to speak of these tones not only in terms of their frequencies but also in such other terms as are employed by musicians and physicists.

In music, tones or notes are arranged in a "musical scale" in the order of their frequencies and in accordance with very definite laws. The method most used in designating the various notes is by the position given them on the lines and spaces of the musical staff. Each note or each line and space on the staff is also referred to by one of the seven letters *a, b, c, d, e, f* and *g.* These letters are applied in this order as one proceeds up the

[1] Fletcher, Harvey, The Physical Criterion for Determining the Pitch of a Musical Tone, *Physical Review,* March, 1924.

scale, each succeeding series of seven notes being correspondingly lettered. The note known as "middle *c*" occupies the single line between the base and the treble staffs, as shown in Fig. 53. This establishes the relative positions of all the others above and below it. These notes and designating letters on the staff correspond, for concrete instance, to those of the white keys on the piano. So in the case of any musical instrument, a definite line or space position on the staff indicates a particular note on the instrument.

FIG. 53.—Musical notation.

Each group of seven notes on the staff, together with the first note in the next group above, cover a musical interval known as an "octave." According to the system of notation just described, therefore, the succeeding notes marking octave intervals will always be indicated by the same letter. Thus an octave extends from one *c* to the next *c* above or below it and, likewise, for any other letter. An octave may begin on any note

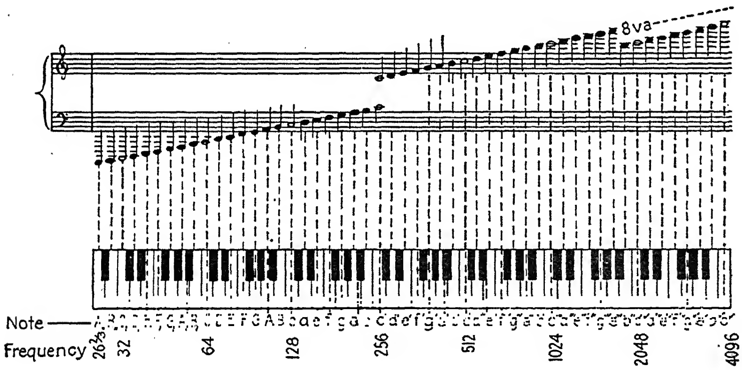


FIG. 54.—Notation and frequency of tones on piano.

and will include the next seven notes above or below it. Thus any consecutive group of eight notes such as *c, d, e, f, g, a, b, c* or *g, a, b, c, d, e, f, g* covers an octave.

The position of any note on the staff is in itself a complete identification of the note, but when it is desired to refer to a particular detached note without recourse to the staff, it is necessary to adopt some system of marking to distinguish the different notes of the same name from each other. There are a number of such systems of notations, some of them so similar as to be likely

to be confused among themselves. The letter notation indicated in Fig. 54 is the one introduced by Helmholtz and the one perhaps most employed by physicists today. For convenience, this notation is associated in this figure with the piano keyboard as well as the musical staffs. In this the first seven notes of the octave ending with middle c are designated merely by small letters, c, d, e, f, g, a and b. For the successive octaves above this the same small letters are used but distinguished from each other by exponential markings, such as, c^{I}, c^{II}, c^{III}, c^{IV}, etc., or c^1, c^2, c^3, c^4, etc. Similarly, the successive octaves going down the scale below the octave c-c^{I} are designated by corresponding capital letters, like letters in this case being distinguished by numerical subscripts. Thus, the successive c's covering the entire range of the piano become: C_1, C, c, c^1, c^2, c^3, c^4 and c^5.

The position of a note on the musical scale is determined by pitch, and pitch only. As pitch in turn is dependent on frequency alone, it is evident that each note must represent a particular frequency. Here again there is some divergence of practice. From time to time there have been various standards of pitch, some of which have been set by international conference. The one most employed by physicists, and shown in Fig. 54, fixes middle c, which is c^1 on the Helmholtz notation, at a pitch corresponding to 256 cycles per second. This determines the pitch of all other notes on the scale, according to the relationships now to be briefly referred to.

The musical scale, as it has been developed by occidental peoples through a long period of time, has as one of its fundamental characteristics the property that in ascending the scale from any note, the frequency doubles as each succeeding octave of that note is reached, and, conversely, in descending the frequency is halved at each octave. Knowing that the frequency of middle c (c^1) is 256, this at once establishes the frequencies of all other c's in the scale. Thus:

Notation......	C_3	C_2	C_1	C	c	c^{I}	c^{II}	c^{III}	c^{IV}	c^{V}	c^{VI}
Frequency....	8	16	32	64	128	256	512	1,024	2,048	4,096	8,192

Having established the relationship of the octave intervals, we may inquire into that of the notes within the octave. In striking two notes on the piano simultaneously, even a person

without musical training will have no difficulty in recognizing that certain combinations produce pleasing effects. These combinations are termed *concords.* As opposed to these, other combinations produce relatively disagreeable effects and are called *discords.* The closest concord is that of the octave in which, as we have seen, the ratio of frequencies is always as 2:1.

The concords within the octave c, d, e, f, g, a, b, c^1, together with the frequency ratios of the higher to the lower tones are as follows:

Interval of concord	Lower tone	Higher tone	Frequency ratio
Octave	c	c^1	2:1
Fifth	c	g	3:2
Fourth	c	f	4:3
Major third	c	e	5:4
Minor sixth	e	c^1	8:5
Minor third	a	c^1	6:5
Major sixth	c	a	5:3

Knowing the frequency of c^{I} to be 256 cycles, this table of concords at once determines the frequencies of five other notes, c, e, f, g and a, within the octave. Those of the two remaining notes, b and d, may be similarly determined by considering another octave beginning with g. In this octave g, a, b, c^{I}, d^{I}, e^{I}, f^{I}, g^{I} the major third will be g-b, and the major fifth g-d. Applying the same ratios for these concords as in the previous case, we arrive at the ratio of 15:8 for the interval c-b, and of 9:8 for the interval c-d. We have then the following frequencies and their ratios within the octave c — c^{I}:

Note	c	d	e	f	g	a	b	c^{I}
Ratios	1	$9/8$	$5/4$	$4/3$	$3/2$	$5/3$	$15/8$	2
Frequencies	128	144	160	$170\frac{2}{3}$	192	$213\frac{1}{3}$	240	256

The study of why these intervals exist in our musical scale would be an interesting one, but it would lead us far afield. Suffice it to say that like the grammar of language, the *diatonic scale,* as it is called, has been a slow evolution by natural selection and, having been thus evolved, the rules governing the relationship have been found afterward.

The note on which an octave begins determines the *key* of the scale. The octave just considered is, therefore, in the key of *c*. For all other octaves in this key the same ratios would apply. In music, however, it is necessary to use various keys; that is, it must be possible to consider the octave as beginning on any note. In attempting to construct the diatonic scale beginning on other key notes, it is found that the notes so far considered, that is, those represented merely by the lines and spaces on the staff or by the white keys on the piano, are not sufficient. Others must be supplied to fill certain intermediate intervals, and for this purpose the various sharps and flats, as represented by the black keys on the piano, for instance, are supplied.

It is also found that when the correct ratios of the diatonic scale are applied for one key, as just described for the key of *c*, the ratios between the various notes do not quite work out in the proper manner when the octave is considered as beginning on some other key note. This makes necessary a slight departure from the intervals of the diatonic scale for such instruments as the piano, where the strings are necessarily tuned to fixed frequencies and where the exact application of the diatonic scale could be made for any one key but would result in certain departures from the proper frequency ratios for any other key. In order to meet this difficulty the so-called "tempered" scale is necessary for all fixed tone instruments, if they are to be played in various keys. The tempered scale, however arrived at, is naturally a compromise between two courses: one, of having an instrument tuned perfectly in one key and badly in all the others, and, the other, of having an instrument tuned slightly but equally imperfectly for all keys. For such instruments as the violin and the voice no such compromise is necessary.

While the standard pitch generally employed by engineers and physicists, as illustrated in Fig. 54, gives middle *c* a frequency of 256 cycles, a slightly higher pitch is now largely employed in music. In the International Tempered Scale a^{I}, the first *a* above middle *c*, is given a frequency of 435 cycles. This is taken as the *key note* from which all others are tuned. The selection of this particular note is probably due to the fact that it is the note from which the violins in an orchestra are given their pitch for tuning the open *a* string. This gives middle *c* a frequency of 258.65 cycles instead of 256 on the older standard scale.

We may conveniently refer to the piano to obtain concrete ideas of the pitch of tones within the range most often encountered in everyday life. On a grand piano, of so-called "$7\frac{1}{3}$ octaves," the frequency range from the note A_2 at the lower end of the keyboard to the note c^{V} at the upper end is from 27.19 to 4,138.44 cycles, if the piano is tuned according to the International Tempered Scale. With this tuning, as stated, the frequency of a^{I} is 435 cycles. The frequencies of any other *a*'s or *c*'s may be readily determined by doubling for the octaves immediately above and halving for those below.

These odd figures of the International Scale are, however, difficult to hold in mind. Those of the scale in which middle *c* has a frequency of 256 (Fig. 54) are sufficiently close to the ordinary piano tuning for most practical purposes and are easily remembered because all of the *c* notes are even powers of 2.

This range represented by the piano keyboard extending from frequencies of about 27 to those slightly above 4,000 cycles represents about the range of fundamental frequencies ordinarily employed in music. The principal exception to this statement is to be found in the tones of the lowest organ pipes. In these tones having frequencies as low as 16 cycles per second (C_3) are usually found, and in some the range extends an octave below this to C_4, with a frequency of 8 cycles. Such an organ pipe has an air column 32 feet long. Its tone, if such it may be called, would be appreciated by many ears as a series of separate pulsations and the sensation produced would be nearer that of feeling than of hearing. Such a tone would be said, for these ears, to lie outside the "area of audition" or beyond the "threshold of hearing." Whether properly classifiable as tone or not, these extremely low rates of vibration seem to have value in organ music as furnishing a foundation upon which to build the higher tonal structure.

We must clearly distinguish between pitch and vibrational frequency in speaking, for instance, of the range of frequencies in the sounds of a musical instrument. In all ordinary tones pitch is determined by the fundamental tone corresponding to the lowest frequency in that tone. Many other higher frequencies may be present in the tone, but they have no effect on the pitch. Thus, while the range of pitch represented by the piano keyboard is, with the exceptions just noted, about that

ordinarily employed in music, the actual range of frequencies occurring in musical tones will extend much higher.

There has been much conflicting data as to the limits of audible frequencies. This has been caused in part by rather wide differences in the acuity of different ears, in part by the fact that the ability of the ear to sense sounds of very low or very high pitch depends somewhat on the loudness of the sound and also in part by imperfections in the methods of observation. The most modern determinations of the range of pitch, which a normal ear is capable of sensing, extends from a frequency of about 20 cycles as the lower limit up to about 20,000 cycles as the upper limit, a range of about ten octaves. Fortunately however, as will be pointed out in later chapters, in the electrical transmission of sound we are not required to make use of nearly so wide a range.

Loudness is the second of the three characteristics by which tones are differentiated from each other. The loudness of a tone depends, in general, on the energy of vibration; but it must be remembered that loudness as measured by the ear corresponds to the degree of sensation produced, while the energy of vibration is a physical attribute of the sound waves producing the sensation. The one is psychological the other physical. It is not to be expected, therefore, that loudness will be exactly proportional to the energy of the sound waves. Even in the same ear the sensation of loudness is not exactly proportional to the energy of the sound waves throughout the entire range of frequency. For different ears there is a more marked variation. Obviously, a given train of sound waves will produce a much louder sensation in a normal ear than in one that is partially deaf. It may be said, however, that for normal ears and throughout the most useful range of pitch, the loudness of a sound may be considered as roughly proportional to the energy of the sound waves.

The energy of a train of sound waves varies as the square of the amplitude[1] and also as the square of the frequency. For tones of the same pitch, therefore, loudness will vary roughly as the square of the amplitude of vibration. For tones of different pitch but of the same amplitude, the loudness will vary roughly as the square of the frequency. For tones varying both in

[1] Square-root-of-the-mean-square value of amplitude, or simply root-mean-square value, ordinarily abbreviated r.m.s.

frequency and amplitude, the loudness will vary roughly as the square of the product of the two.

In Fig. 55, Dayton C. Miller has shown the wave forms of several simple tones having various degrees of loudness. Of the notations on this figure, n represents the frequency, A the amplitude and I the intensity of sensation or the loudness. In the upper curve a the frequency and the amplitude are each assumed to have a value of unity, and, since the loudness is proportional to the product of the squares of the two, its value may also be considered as unity. In curve b the frequency is

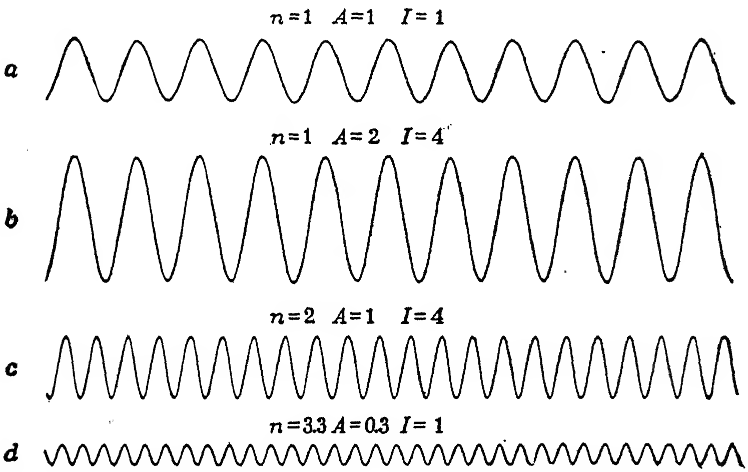


FIG. 55.—Curves representing simple sounds of various degrees of loudness.[1]

the same, but the amplitude is twice as great, hence the loudness is four times as great. In curve c the frequency is doubled, the amplitude remaining the same as in the first curve, and again the loudness is increased fourfold. In the lower curve d the frequency and amplitude have been so chosen as to make their product substantially unity, again resulting in a loudness of unity. The sounds of waves a and d will thus be of equal loudness. Those of b and c will also be alike in loudness, but four times as loud as those of a and b.

When the tones are complex the treatment of loudness is not quite so simple. In Fig. 56, taken from the same source, there is shown a lower complex wave c compounded of the two simple

[1] Reprinted from MILLER, D. C., "The Science of Musical Sounds," The MacMillan Company.

waves a and b above it. The amplitude, frequency and loudness of the two simple waves are as indicated and the loudness of the complex tone is the sum of that of the components. This brings out in striking manner the fact that the harmonics of a complex tone may carry, contrary to the popular conception, a far greater amount of energy than the fundamental wave, and, as a result, contribute far more to the loudness sensation than the fundamental. In the case of Fig. 56 the amplitudes of waves a and c are not greatly different; the pitch of the two sounds would be the same because each has the same fundamental frequency;

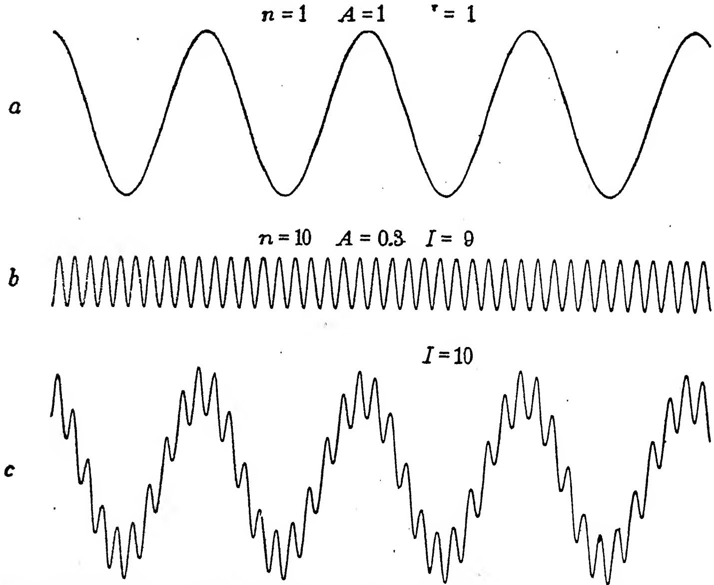


Fig. 56.—Relative loudness of two simple sounds and their combination.[1]

yet sound c would be ten times as loud as sound a because of the harmonic wave b carried on its fundamental.

As will be shown, it is of great importance in the telephonic transmission of speech that the sounds at the receiving instrument shall be loud enough to be heard and interpreted easily. It is the lower limit of audibility that is of most importance in the design of telephone transmission systems. On the other hand, the received sounds must not be so loud as to cause annoyance or

[1] Reprinted from Miller, D. C., "The Science of Musical Sounds," The MacMillan Company.

pain or to detract from their intelligibility. Between these two extremes there is a wide range of permissable loudness, in which the transmitted speech, unless lacking in other respects, may be easily heard and interpreted. In varying the loudness, however, it is of paramount importance, as affecting intelligibility, that the *relative amplitudes* of the component frequencies be preserved as far as possible. The variations in amplitude should affect all frequencies in the same proportion. This, however, bears directly on the quality or timbre of sounds, the characteristic next to be considered.

Quality or Timbre.—This is the third distinguishing characteristic of sound. It is a matter of ordinary observation that sounds may be alike in pitch and also in loudness and yet produce quite different effects upon the ear. The differences between the tones of a piano, flute, violin, oboe and the human voice, for instance, all sounded at the same pitch and all with the same loudness, are those of quality alone. Being of the same pitch, they all have the same fundamental simple harmonic vibration. The difference lies in the number of harmonics present and in their relative intensities.

It is often stated somewhat loosely in textbooks on sound that quality of tone depends on the wave form. In many cases this is true, but there are important exceptions. Moreover, it seems to be a rather illogical way of looking at the matter, for, in so far as we understand the workings of the ear, it does not function at all as a *form* analyzer but as a *frequency* analyzer. In our reasoning about sound we find it convenient to consider wave form, but the ear needs no such roundabout method. It interprets frequencies directly. We have only to refer again to Figs. 43 and 44 to see the fallacy of holding wave form responsible for quality. The two resultant curves of those figures are quite different in form; yet, having the same component simple waves, they would produce exactly the same tone. Here the difference in resultant wave form is due solely to the different phase relations of the component simple waves, and experience shows that *changes in phase produce no noticeable change in the character of the sound heard.*

We may conclude, therefore, that quality or timbre is varied by changing the number of components or by changing their relative amplitudes but not by changing their phase relationship.

CHAPTER VI

THE SENSATION OF SOUND

An understanding of the sensation of sound and of the structure and action of the hearing mechanism is of scarcely less importance than that of the physical vibrations which cause the sensation.

Passing through the spectrum of sound vibrations from the lowest to the highest frequencies, it is found that the ear will respond only to vibrations within a certain range of frequencies. Outside this audible range the vibrations are inaudible, just as in the spectrum of light vibrations there are visible and invisible ranges. It is found also that the degree of loudness has a marked effect on the range of frequencies which the ear is able to perceive, sounds both above and below a certain critical loudness showing a marked falling off in the length of the audible frequency spectrum. Again, the sensitiveness of the ear is not uniform throughout the length of the audible spectrum.

These and other characteristics make it important to gain at least a fundamental understanding of the ear mechanism and the sense of hearing. In telephony the commodity transmitted is sound, principally the sound of speech. Of this commodity the ear is the ultimate and sole consumer. Obviously, ordinary commercial considerations demand that we should cater to the peculiarities of this consumer. Money spent in delivering to the ear sound waves of a kind which it has no power to perceive, or little power to digest or use, is likely to be money wasted. Much is to be gained, therefore, by a study of the sense of hearing to ascertain what kind of commodity, in the form of sound waves, the ear will find most useful in the reception and interpretation of articulate speech or, in the less important function, the appreciation and enjoyment of music.

Figures 57, 58 and 59 are helpful in obtaining a preliminary understanding of the mechanical structure of the organs of hearing.[1] Figure 57 shows the general relationship of the outer

[1] For the photographs from which these were made I am indebted to the American Telephone and Telegraph Company.

and inner ears with respect to the bony structure of the skull. Figure 58 (left) is a sectional view through the skull bone looking toward the outer ear and shows the small chamber of the middle

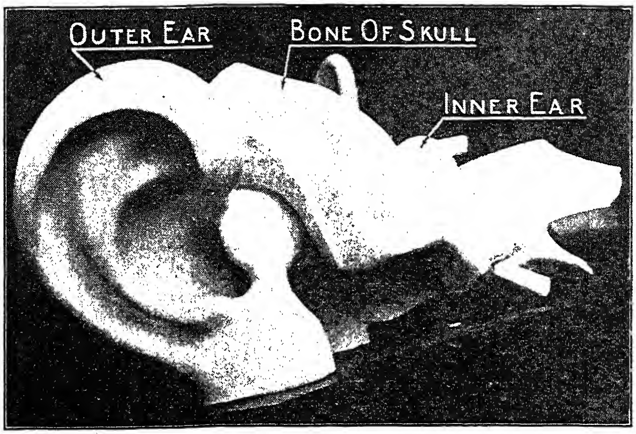


FIG. 57.—Outer and inner ear with reference to bony structure of skull.

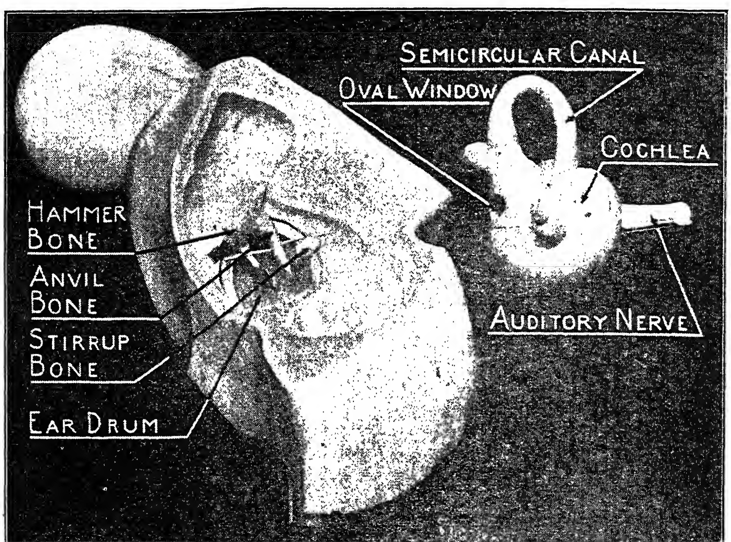


FIG. 58.—Left, cavity of middle ear. Right, bony structure of inner ear.

ear. The ear drum or tympanic membrane shown in this view is tightly stretched across the ear canal leading from the outer ear. This membrane receives the sound vibrations entering the

outer ear, and, by means of the chain of three small bones, called, respectively, "hammer," "anvil" and "stirrup," passes them mechanically through the middle-ear chamber to the oval window which forms the entrance to the cochlea or inner ear. Figure 58 (right) is an external view of the bony structure of the inner ear. It shows the oval window through which the sound vibrations are received from the small bones or "ossicles" of the middle ear and, at the right, the auditory nerve, the "cable" which connects the inner ear with the brain.

Figure 59 shows the inner ear partly in section. It is within the spiral chamber or cochlea that the sound vibrations act on the auditory nerve. The three semicircular canals shown in

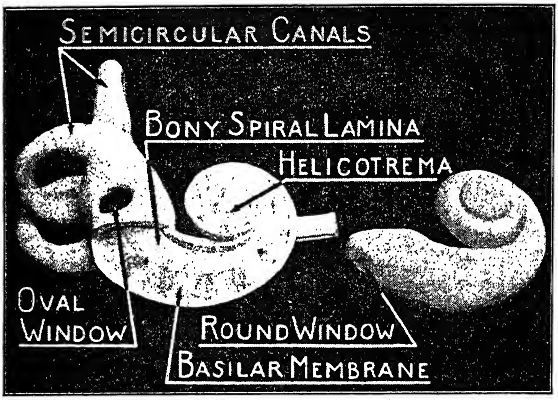


FIG. 59.—Inner ear partly in section.

Figs. 58 and 59 may be passed with the comment that they probably have nothing to do with the sense of hearing. Their functioning is not fully understood, but it is now quite definitely known that they are in some way associated with our sense of balance, by means of which we maintain our equilibrium in walking or in other bodily action.

The spiral chamber of the cochlea is filled with a fluid and is divided throughout its length by a long, narrow flexible membrane called the "basilar" membrane. There are, thus, within the cochlea two long narrow chambers separated throughout their lengths by the basilar membrane; these being rolled up into spiral formation, much like that of a conch shell. At their extreme inner ends there is an opening between the two chambers called the "helicotrema." One of the chambers, at its larger

outer end, connects with the chamber of the middle ear through the oval window, while the other chamber, in similar manner, connects with the middle-ear chamber through the round window.

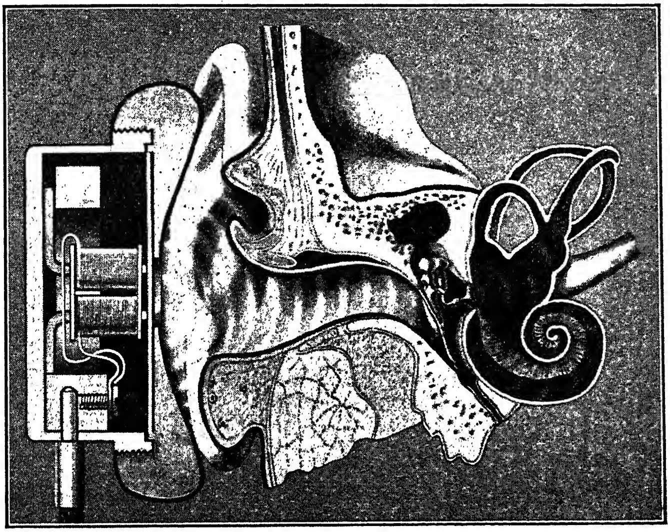


FIG. 60.—Cross-section of ear and telephone receiver. (*Courtesy of Bell Telephone Laboratories.*)

Figure 60 gives a more comprehensive cross-section view of the ear structure and also shows the relationship with the outer, middle and inner ears of the working parts of a telephone receiver when held in proper position against the outer ear.

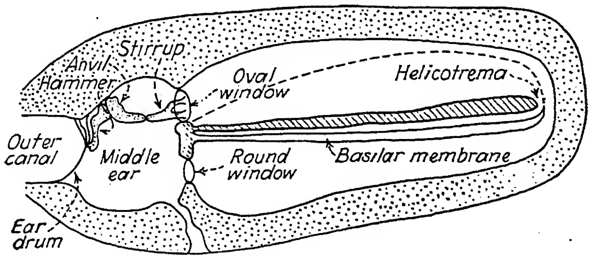


FIG. 61.—Diagram of auditory function.

The various parts of the organs of hearing, with their connecting openings and passages, are shown diagrammatically in Fig. 61. In this the cochlea or inner ear is shown uncoiled or rolled out into a plane surface. It is to be noted that there is a continuous

passage through the fluid of the cochlea between the two windows communicating with the middle ear. This passage is traced from the oval window, through which sound waves are delivered to the cochlea, thence through the spiral convolutions on one side of the basilar membrane to the helicotrema and thence through the spiral convolutions on the other side of the basilar membrane to the round window which again communicates with the middle-ear chamber.

The basilar membrane is approximately 1¼ inches long (more accurately 31 millimeters), and throughout its length from the oval window to the helicotrema the cells of the auditory nerve are distributed. There are about 4,000 of these nerve cells, each having four or five fiber hairs, making in all from 16,000 to 18,000 terminal units distributed along the basilar membrane.

In spite of a vast amount of investigation by eminent scientists, there is still much difference of opinion about the method of functioning of the inner ear. It is clear enough how the sound waves impinging on the ear drum through the outer ear passage are transmitted by the ossicles of the middle ear and the membrane covering the oval window to the fluid of the inner ear. From this point on theories differ. Two of the principal ones may be briefly mentioned:

In the theory advanced by Helmholtz it was held that a series of sound resonators, each sharply attuned to a given frequency, were progressively arranged along the length of the basilar membrane. Each of these resonators would, therefore, respond to a single pitch and stimulate the corresponding nerve terminus accordingly.

The other of the two theories to be referred to here is the so-called "maximum amplitude theory." This was originally formulated by Gray in 1899, and a modification of it is now being put forward by Fletcher[1] as probably the one most nearly in accordance with the mass of experimental facts recently developed by himself and his associates, Messrs. Wegel, Lane and others.

Briefly, this theory, as stated by Fletcher, is that the entire basilar membrane vibrates for every incident tone, and that for

[1] FLETCHER, HARVEY, Physical Measurements of Audition and Their Bearing on the Theory of Hearing, *Journal of the Franklin Institute*, September, 1923.

each frequency there is a corresponding spot on the membrane where the amplitude of vibration is greatest. The theory assumes that only those nerves are stimulated which are at the particular parts of the membrane that are vibrating with more than a certain critical amplitude; and that the brain judges the pitch from the part of the membrane where the nerves are stimulated.

If it is assumed that there is no yielding of the cochlea walls under varying pressure, then any movement imparted to the liquid of the cochlea by the stirrup bone acting through the oval window will cause an equal and opposite movement of the liquid at the membrane covering the round window. This transfer of movement is, of course, due to the mass movement of the liquid, and this can occur from one cochlea chamber to the other either by a deflection of the basilar membrane or by an actual flow through the helicotrema.

Remembering that the response of the mass of liquid increases rapidly as the velocity of movement decreases, it is clear that for very slow movements imparted by the stirrup at the oval window there will be a bodily movement of the liquid through the helicotrema, the whole transfer, in this case, being effected without deflecting the basilar membrane at all. Hence it is that for very slow frequencies, from zero up to about fifteen or twenty cycles per second, the basilar membrane is not perceptibly deflected and no sensation of sound results because none of the nerve termini are stimulated.

On the other hand, for very high frequencies the effective inertia or "mechanical impedance" of all the moving parts, from the eardrum in, is enormously increased. In fact, the effective inertia of the small bones of the inner ear may be so great for these frequencies as to absorb all of the impinging sound energy, leaving none to cause appreciable movement of the cochlea fluid at the oval window. This clearly accounts for the fact that vibrations of 20,000 cycles per second and above cause no disturbance of the basilar membrane and excite no sensation of sound. The ranges of the acoustic spectrum above and below the limits of audition are thus accounted for. The lower frequency vibrations pass around the basilar membrane without causing it to move, while those of very high frequency have all their energy absorbed by the mechanism of the middle ear before reaching the oval window.

Consider, now, how the maximum amplitude theory accounts for the action of the inner ear throughout the audible range of vibration. We have seen that except for a slight yielding of the walls of the cochlea, any mass movement of the fluid, imparted through the oval window, can take place only by flow through the helicotrema or by displacement of the basilar membrane. While the entire basilar membrane may partake of this vibration, a consideration of the dynamic features of the cochlea mechanism will show that for each frequency there will be a well-defined maximum displacement at some definite point along the length of the membrane. For the lower frequencies involving slower movements the inertia friction and elasticity of the parts will result in the movement of comparatively large masses of the fluid through relatively long paths, thus bringing the point of maximum amplitude of vibration toward the remote end of the basilar membrane near the helicotrema.

As the frequency of vibration imparted by the stirrup bone at the oval window increases the dynamic reactions within the cochlea will result in a smaller mass of the fluid and a shorter path through it being brought into play, resulting in the point of maximum amplitude being shifted along the basilar membrane toward the oval window. Finally, for very rapid vibrations that are still within the audible range, the dynamic conditions within the cochlea will be satisfied only by a movement of a very small mass of the fluid through a very short path, with the result that the point of maximum amplitude of vibration of the basilar membrane will be at the end near the oval window.

To express this theory of the selective action of the inner ear in a somewhat more concise way: It is evident that there are an infinite number of paths through the fluid of the cochlea from the oval window where the pressure variations are received to the round window where they are relieved. These paths vary in length, the longest involving a loop reaching to the helicotrema and back, the shorter ones successively taking shorter cuts through the basilar membrane at points nearer and nearer to the oval window. As these paths will successively involve smaller and smaller mass movements of the cochlea fluid and also will have successively varying other dynamic constants, such as frictional resistance and elasticity, it is quite clear that each of them will be dynamically resonant to a different frequency of vibrations. The lower frequencies will naturally

choose the longer paths and the higher frequencies the shorter. Tones of lowest pitch will, therefore, affect the basilar membrane at the helicotrema end, those of highest pitch at the oval window end, others arranging themselves along the membrane in the order of their frequencies.

Thus, while the maximum amplitude theory of hearing does involve resonance, it is quite different from the Hemholtz theory which contemplated, in effect, several thousand small sound resonators distributed along the basilar membrane, each attuned to a different frequency from all the others. One difficulty with the Helmholtz theory is that it does not seem to conform to the actual dynamic possibilities of the ear structure, which, of course, are now much better known than when Helmholtz formulated his theory.

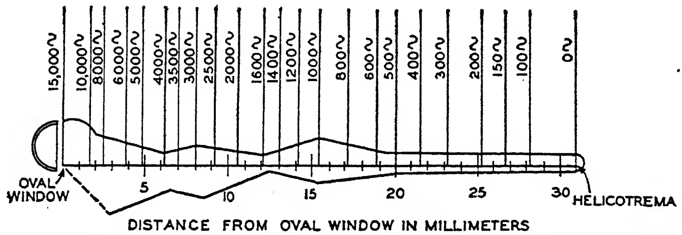


FIG. 62.—Characteristic frequency regions on basilar membrane. (*Courtesy of Bell Telephone Laboratories.*)

The various positions on the basilar membrane of maximum response to different frequencies throughout the audible range, as calculated by Wegel and Lane, are shown in Fig. 62. This shows the distribution between limiting frequencies of zero and 15,000 cycles, but it is pointed out that had these limiting frequencies been considerably varied, as from 50 to 20,000 cycles, it would have made practically no difference with the distribution along the intermediate portion of the scale, because these ranges adjacent the upper and lower limits are of negligible length as compared with what may be considered the real working range between them.

Anatomical examination has shown that the distribution of the auditory nerve terminals along the length of the basilar membrane is quite uniform. This makes possible an interesting deduction which helps to illustrate the astounding sensitiveness of this membrane. The smallest difference in frequency which the ear is just capable of detecting is known, as is the length of

the basilar membrane and the distribution of the points along it responsive to different frequencies (Fig. 62). With this data one can, by a brief calculation, show that the ear can recognize, as distinct, two frequencies corresponding to points on the basilar membrane as close together as 0.02 millimeter, less than 1/1,000 inch. Comparing this with the sensitiveness of other parts of the body, the minimum distance between two pin points that can be recognized as separate on the back of the hand is about 1¼ inches and on the finger tips about 1/10 inch.

The function of the auditory sense, according to Wegel,[1] "is to detect sounds of various kinds and wave shapes varying over a range of pressure on the ear drum from about 0.001 dyne to 1,000 dynes, and over a considerable part of this range to differentiate with certainty between complex sounds so nearly alike that no existing physical apparatus can separate them."

In the foregoing discussion of the action of hearing we have seen something of the marvelous sensitiveness of the ear mechanism in the determination of pitch, but, from the statement just quoted, it appears that the ear must include in its powers the ability to discriminate between vibrations on the ear drum through a range in pressure changes in which the maximum is something like a million times as great as the minimum change just discernible, and through a range of energy change in which the maximum is in the order of a million million times the minimum. To give an idea of the minimum value of the pressure change which the ear can sense, it is pointed out that a pressure change of 0.001 dyne on the eardrum corresponds to a change of about one billionth of the ordinary atmospheric pressure.

In the theory just discussed the brain is assumed to determine the pitch of a sound by the point along the length of the basilar membrane at which the peak of vibration amplitude occurs. It is quite obvious that it associates the loudness or intensity of the sound in some way with the degree of violence of agitation of the basilar membrane in the region of that peak.

It may be that the degree of loudness is determined by the amplitude at which the peak point is vibrating under the influence of the sound, the brain discriminating in this respect by the violence of the agitation of the nerve centers at this point. Or, it may be, that the peak is made wider with increased

[1] Wegel, R. L., The Physical Examination of Hearing and Binaural Aids for the Deaf, National Academy of Sciences, *Proceedings*, July, 1922.

amplitude and that the brain senses increased loudness by the wider band of nerve centers thus brought into preceptible action. Perhaps both of these actions enter into the perception of relative sound intensities.

Let us see what happens when we impress on a normal ear drum a simple sound wave of, say, 1,024 cycles, beginning with such slight amplitude as to produce no sensation whatever and gradually increasing its amplitude indefinitely, always holding its frequency constant. At first the pressure changes between the successive crests and valleys of the wave will be so slight as to lie below the lower limit of audibility. As the amplitude is increased, however, to a point where the root-mean-square pressure changes on the drum amount to perhaps 0.001 dyne, a faint pure tone with a perfectly recognizable pitch corresponding to c^3 on the musical scale will be heard. This marks the lower limit of audibility for a tone of that pitch, and the point is referred to as the *threshold of audibiltiy* for that pitch.

The amplitude may then gradually be increased through a long range in which a pure tone of ever increasing loudness but of constant pitch, c^3, will be heard. Finally, probably, when the root-mean-square pressure change amounts to over 5,000 dynes (over 5,000,000 times its value at the threshold of audibility), a change will occur. The sensation of sound and of tone will give way to one of feeling, a tickling sensation which changes to pain if the vibrations are still further increased. The upper limit of audibility as to loudness has been reached, and this point is called the *threshold of feeling* for that pitch.

For each separate pitch throughout the entire audible range of frequency there will be found, as the amplitude of the wave is increased from zero, first an inaudible range, then, successively, a point marking the threshold of audibility, a long audible range of constantly increasing loudness and, finally, a point marking the threshold of feeling where the sense of touch supplants that of hearing.

Figure 63 shows a mapping out of these threshold points throughout the entire range of audible frequencies. Each vertical line represents a given frequency or pitch; each horizontal line corresponds to a given pressure variation in dynes per square centimeter or pressure amplitude, and, therefore, represents a given energy level. As the intensity or loudness of a sound is closely related to the energy, each horizontal line may also be

taken as representing a certain loudness level. The lower curve is made by the points marking the lower curve limit of audibility for each successive frequency, and the upper curve, likewise, by the points where hearing ends and feeling begins.

These curves, so far as they are marked by full lines were made by actual measurements on a large number of normal ears. The dotted portions of the curves were extended by extrapolation, actual observation being difficult in the corresponding upper and lower frequency ranges.

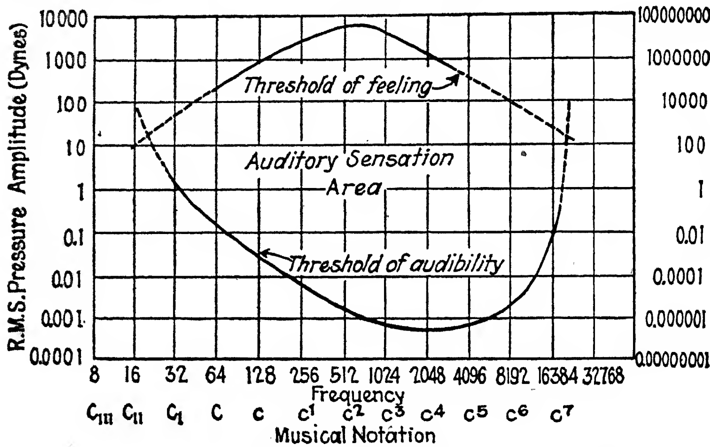


FIG. 63.—Audible ranges of amplitudes and frequencies.

An examination of the lower audibility curve will show that for frequencies below about 500 cycles and above about 6,000 cycles, the sensitivity of the ear decreases quite rapidly, requiring larger minimum pressure variation to become audible. Likewise, for the upper curve the threshold of feeling is encountered at lower and lower intensity levels, as the frequencies diminish below or increase above the ordinary range. These facts cause the minimum curve to bend up and the maximum curve to bend down, thus enclosing an area between them. This is the area of auditory sensation. Obviously, any sound is audible if its pitch and intensity correspond to a point within this area and inaudible if without.

The recent mapping out and exploration of this audible area is an important contribution to the science of hearing. Among other things they throw a great deal of light on the heretofore much debated subject of the upper and lower limits of the fre-

quency range at which sound can be sensed. They show clearly that the pitch limits of audible sound are largely dependent upon the intensity at which the sound is heard. To illustrate: if the sound is impressed on the ear drum at an intensity corresponding to the 10-dyne line of Fig. 63, the lower audible frequency limit would be approximately 20 cycles and the upper about 20,000 cycles. As the intensity is either raised or lowered from the level of the 10-dyne line the ranges of audible frequency are markedly shortened. As an example, if the pressure amplitude is lowered to the 0.1-dyne level, the lowest audible frequency becomes somewhat above 64 cycles and the highest in the neighborhood of 16,000 cycles. If lowered to the 0.001-dyne level, these limiting frequencies become about 800 and 6,000 cycles, respectively. One cannot, therefore, with safety speak definitely of the limiting frequencies of sound sensation without a knowledge of the intensity of the sound vibrations reaching the ear.

The subject of the choice of a practical working unit for loudness is occupying the minds of scientists working in this field. It is obviously important that eventually some usable reference scale shall be adopted to measure and compare the loudness of sounds. In the telephone business, for instance, the loudness of transmission delivered along various points of the line and at its end is the characteristic most used in determining relative efficiencies of transmission. This must not be taken as indicating that a satisfactory degree of loudness will mean satisfactory telephone transmission, but that, for transmission which is otherwise satisfactory, relative degrees of loudness are fairly indicative of relative efficiencies. Transmitted speech sounds may, for instance, be quite satisfactory as regards quality, but the loudness level may be so low as to make hearing them and interpreting them either difficult or impossible. Again, in all of the branches of acoustics a standard of loudness would be useful, if for no other reason than to enable workers in different fields to understand each other's language.

The common measure of the degree of sensation in sound is loudness. As has been shown, while perhaps not directly proportional to the energy of the sound vibrations, loudness is closely related to it. In view of the enormous range of energy ratios of a given audible tone and also on account of the geometrical relationship between sensations and stimuli shown by

psychological research, it is apparent that the scale of loudness ·ay most conveniently and properly be logarithmic. It will e noted that in Fig. 63 the logarithmic scale was used both for the ordinates representing energy ratios and for the abscissæ representing frequency. The rather obvious reason for adopting the logarithmic scale for pitch will be referred to later under the discussion of pitch units.

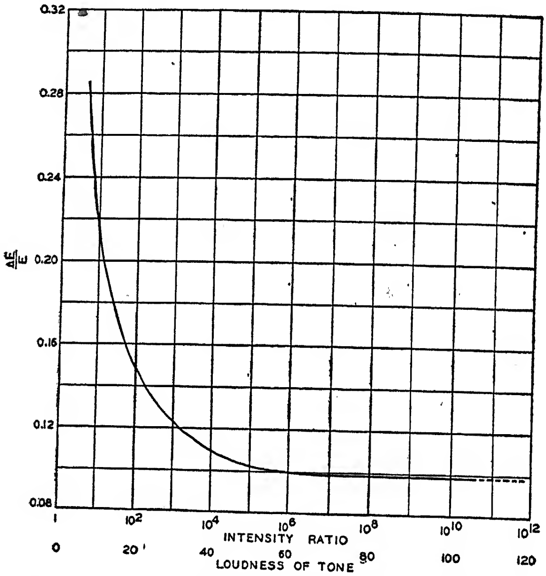


FIG. 64.—Minimum perceptible difference in intensity. (*Courtesy of Bell Telephone Laboratories.*)

We may approach the choice of a loudness unit by considering the minimum change in the loudness of a tone which the ear is just capable of sensing. Figure 64, taken from Fletcher's paper[1] and based on the work of V. O. Knudsen, shows the minimum perceptible differences in intensity or loudness throughout the audible range. In this figure points along the ordinates represent the fraction by which the sound energy must be increased or diminished in order that the change may be just perceptible. The abscissæ are ratios of intensity of tone to intensity at the threshold of audibility for that tone. Thus,

[1] FLETCHER, HARVEY, Physical Measurements of Audition and Their Bearing on the Theory of Hearing, *Journal of the Franklin Institute*, September, 1923.

for a tone whose intensity is 10 times the intensity of that tone at the threshold, a change of 22 per cent in its intensity will be required to produce a perceptible change in loudness. This ratio gradually decreases as the sounds become more intense, *i.e.*, louder, and for sounds having intensities of more than 10^4 or 10,000 times the threshold value, the minimum noticeable fractional change is nearly constant at about one-tenth. This fractional increase required for perceptible change is called the Fechner ratio, and has approximately the same range of values for all frequencies.

The exponents (logarithms to base 10) used in expressing the intensity ratios in the abscissæ of Fig. 64, might, in view of the considerations already mentioned, be suggested for use as points on a corresponding scale of loudness. The units of such a scale would be too large for convenience, as an increase in loudness of one unit would represent a tenfold increase in intensity. Fletcher suggests, therefore, dividing the loudness scale into ten times as many divisions, thus making the points on this scale, in each case, equal to ten times the common logarithm of the corresponding intensity ratio. This is the loudness scale shown on Fig. 64. On it one unit is of the order of magnitude of the smallest change in loudness that is ordinarily appreciable by the normal ear. Multiplying the intensity by 10 adds 10 units to the loudness; multiplying the intensity by 100 adds 20 units to the loudness, and so on, multiplying the intensity by 10^n adds $10 \times n$ units to the loudness.

This unit of loudness is being adopted widely by telephone companies in the United States. It is known as the "transmission unit" (*TU*) or, more recently, as the "decibel," after Alexander Graham Bell. Whenever we measure loudness a comparison is either expressed or implied. We say one sound is louder than another, expressing the comparison, or, if we merely speak of the loudness of a sound, using the unit described above, the implied comparison is with the loudness of that sound at its threshold value. The difference in loudness between two sounds measured in these units is given by the relation:

$$\text{Decibels or } TUs = 10 \log_{10} \frac{E_2}{E_1},$$

where E_2 is the energy of the sound in question and E_1 the energy of the sound to which the first is compared.

If we wish to compare pressure amplitudes instead of energies, then, since energy in a sound wave is proportional to pressure squared, we have:

$$\text{Decibels or } TU = 10 \log_{10} \frac{E_2}{E_1} = 10 \log_{10} \frac{p_2^2}{p_1^2} = 20 \log_{10} \frac{p_2}{p_1}.$$

Let us examine the relative values of a decibel (*db*) and the minimum perceptible loudness change in two instances. First, let us consider a tone, E_1, near the threshold, of energy, say, 10 times the threshold energy. Let us suppose the energy of this tone is increased by an amount just perceptible, giving tone E_2. Figure 64 tells us that this will require an energy increase of about 22 per cent, or $E_2 = 1.22E_1$. Then the change in loudness as measured in *db* can be calculated thus:

$$db = 10 \log_{10} \frac{1.22E_1}{E_1} = 10 \log_{10} 1.22 = 10 \times 0.086 = 0.86.$$

Hence, at this low intensity about 0.9 *db* is perceptible.

Suppose we take another case where the intensity is high, say 10^8 times the threshold intensity. Let us work the calculation the other way around, increasing the loudness by 1 *db* and calculate the energy change. Then

$$1 = 10 \log_{10} \frac{E_2}{E_1} \text{ or } \log_{10} \frac{E_2}{E_1} = 0.1.$$

Whence

$$\frac{E_2}{E_1} = 1.26,$$

showing that the increase in intensity was 26 per cent. Figure 64 shows us that this is over 2½ times the minimum perceptible change for this loudness, indicating that about 0.4 decibel is noticeable in this range.

From the two foregoing examples we see that, under the conditions represented by the curve of Fig. 64, the minimum perceptible difference in loudness varies over the audible range from about 0.9 to about 0.4 decibel. Since the values of the Fechner ratio shown in the curve were taken under the most favorable conditions for observing changes in loudness, the minimum perceptible difference under practical conditions is considerably larger—enough so to make the decibel a fair approximation to it throughout a large portion of the intensity range. Because it represents a fair approximation to the smallest noticeable change in loudness,

because it affords an easily workable unit for mathematical analysis and because it does not depart radically from an earlier widely used unit,[1] the decibel finally has been chosen as the practical unit of intensity or loudness.

The question of a uniform pitch unit which will measure directly the sensation rather than the cause of it may be approached in the same way; but it would appear that the need for it is not nearly so great as that for a loudness unit. Figure 65,

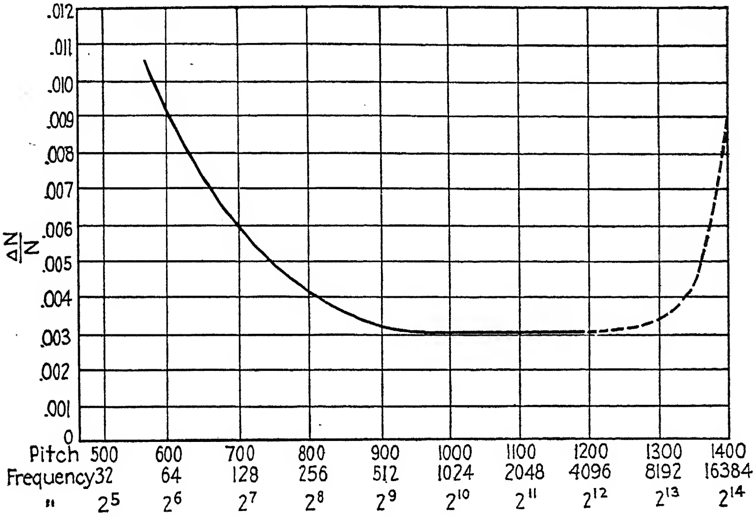


FIG. 65.—Minimum perceptible difference in frequency.

drawn from the same sources as Fig. 64, shows the minimum perceptible differences in frequency throughout the most used portions of the audible range. The two lower scales of abscissas show the frequencies in cycles per second, expressed both in ordinary numbers and in the corresponding powers of 2. The ordinates show the fractional amounts by which the frequency of a tone must be varied in order to produce a just noticeable difference in pitch.

This curve shows that for the lower frequencies the ear is relatively less sensitive to pitch changes than for the most used portion of the upper range. At a frequency of 64 cycles a change of 0.009 in the frequency value is required to produce a noticeable change in pitch. At 256 cycles a change of 0.004 is noticeable,

[1] The loss through a mile of standard cable.

and the ratio becomes practically constant at about 0.003 for a considerable range of frequencies above about 500 cycles.

The use of the logarithmic scale to the base of 2 in laying out frequencies has long been common practice and springs naturally from the fact that each doubling of the frequency marks an octavé in pitch. If the logarithm to the base 2 of the frequency were used directly to express the units along the pitch scale corresponding to frequencies, a change of one unit would mean a change in pitch of just one octave. Of course the ear can detect very much smaller changes than this, and, in order to bring the unit into approximate correspondence with the smallest

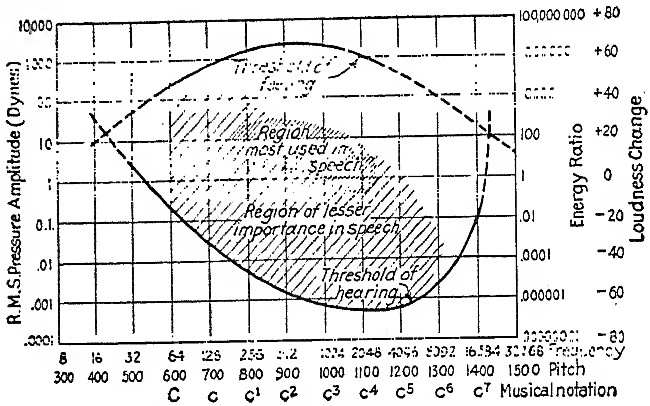


Fig. 66.—Area of auditory sensation.

noticeable change in pitch, Fletcher proposes the adoption of a pitch scale wherein the successive octaves are marked by 100 times the logarithm to base 2 of the frequency. This makes each octave interval in the frequency scale contain somewhere near 100 just perceptible units in the pitch scale. This proposed pitch scale is used for the upper set of abscissa markings in Fig. 65. Care must be taken not to confuse it with frequency.

In Fig. 66 the curves enclosing the auditory sensation area are again given with the just discussed sensation scales of loudness and pitch added. Here, however, the zero point of the loudness scale has been shifted to the 1-dyne amplitude level and the points on the loudness scale are marked positive or negative according as they are above or below this arbitrary reference level. Quantatively they are the same as in Fig. 64 with plus and minus signs regarded.

The consideration of this auditory sensation area diagram in the light of these suggested loudness and pitch sensation scales makes possible an interesting calculation to determine the total number of different pure tones that the normal ear can perceive. If, across the area of sensation of Fig. 66, two vertical lines are ruled just far enough apart that the horizontal distances between them will represent the smallest increment of frequency that is just noticeable as producing a change in pitch, then the narrow vertical area enclosed by these lines and the two threshold curves will contain all the pure sounds that the ear can perceive at that pitch. If this narrow vertical pitch area be ruled off with horizontal lines in such manner that the vertical distances between them will represent the smallest noticeable increment in loudness, each of the small rectangles so formed will represent the same sound sensation, for the ear can detect no difference of pitch or loudness within the dimensions of that area. Obviously, also, each rectangle will represent a sound which the ear can detect as different from all others, for a change into the rectangles either immediately above or below will represent noticeable differences in loudness, and into the area on either side a noticeable difference in pitch. If a similar process were followed for the entire auditory sensation area, the total number of small rectangles so formed would represent the total number of pure sounds that the ear can sense as different.

Fletcher has integrated this area of audible sensation on the basis of the smallest noticeable increments of pitch and loudness as determined experimentally by Knudsen, and his calculations bring him to the conclusion that the ear can distinguish no less than 324,000 different pure sounds.

This total is confined to sounds produced by simple sinusoidal vibrations—pure tones alike in quality. If complex sounds were included in the calculation, so that variations in quality would be sensed, the number of distinguishable sounds would, of course, be enormously greater than this figure. This line of reasoning affords another striking proof of the almost unbelievable sensitiveness of the ear.

It is found that pure tones, regardless of pitch, will appear practically alike in loudness if they are an equal number of loudness units above the threshold intensity value—that is, if they are of the same *absolute loudness* on the scale of units we have chosen. We must, of course, distinguish between

absolute loudness, which is thus measured above threshold value, and loudness measured above any arbitrary intensity level as, for instance, in Fig. 66. With complex sounds the determination and measurement of loudness is not so simple. If two complex sounds are adjusted to such intensity levels as to appear of equal loudness and are then magnified or reduced by equal amounts, they may not appear equally loud at the new level. Hence, we can no longer use the threshold value as the basis for absolute loudness but must choose some arbitrary and easily reproducible standard of comparison. The one adopted by the Bell System in the United States is a pure tone of 700 cycles. The 700-cycle tone is adjusted in intensity until its loudness equals that of the complex tone in question. The loudness of the 700-cycle tone is then easily measured relative to its threshold, and the number of units found is said to be the loudness of the complex tone.

As will be seen on Fig. 66 the pressure amplitude value for the lowest audible tone of 700 cycles is exactly 0.001 dyne. This was one reason for choosing the 700 cycle tone as a standard. Another reason was that the 700 cycle frequency corresponded quite closely to the pitch at which the loudest tones used in conversational speech occur.

Of the enormous ranges through which the ear is capable of sensing variations of both pitch and loudness of sounds, only portions of these ranges are ordinarily used in hearing the sounds of speech and even smaller portions are absolutely necessary. This is fortunate in many respects. For instance, a person afflicted with partial deafness, which results in making inefficient or useless portions of his auditory sensation area, may, perhaps, by using artificial devices, shift the vibrations to some other portions where his sense of hearing is more acute. Again, in telephony, the fact that only a portion of the relatively small frequency range employed in speech will suffice for intelligible transmission and the fact that the ear can with great facility adapt itself to different intensity levels to meet the requirements of the occasion greatly simplify the telephone transmission problem. Far less expensive equipment and plant are necessary than would be required if the whole or large portions of the auditory sensations area had to be covered.

On the plot of the auditory sensation area (Fig. 66) the portion of the region within the area most used in speech is marked by

heavy shading and the region of lesser importance in speech by lighter shading. It is with the smaller region of the darker area that the telephone engineer is most concerned.

Mention has been made in Chap. V of the fact, which is quite contrary to general belief, that the pitch of a complex sound is not necessarily determined by the vibration frequency of the fundamental tone in that sound. Experiment shows that the fundamental frequency and, indeed, several of the lower overtones may be eliminated from the sound without any change whatever in its pitch, though the quality may be somewhat altered. A large amount of experimental data proving this beyond question is given by Fletcher.[1] As an example, the tone of the voice sounding "ah" at the pitch *d* (145 cycles) was uttered into a telephone transmitter forming a part of a practically distortionless telephone system, the receiver of which was employed in observing the sounds. The system was carefully designed to transmit all of the components of the original sound with the same relative intensities and to introduce no new components (hence a distortionless system capable of producing a faithful copy of the original sound). Into this system selective electrical filters could be introduced so that bands of frequencies could be eliminated from the received sound.

The changes resulting from the elimination of various components of this tone are tabulated below:

Effect on the Pitch and Quality of a Musical Sound Caused by the Elimination of Various Components

"Ah" sung at pitch *d*, 145 cycles

Components eliminated	Frequencies eliminated	Pitch	Quality
F	0– 250	No change	Inappreciable change
F and 1–2	0– 500	No change	Small change
F and 1–4	0– 750	No change	Large change
F and 1–7	0–1,250	No change	Very large change
F and 1–9	0–1,500	Uncertain	Noise
6–∞	1,000–∞	No change	Small change
3–∞	500–∞	No change	Large change
F and 1–2 and 6–∞	0–500 and 1,000–∞	No change	Very large change

Note.—In first column, F refers to fundamental and numerals to harmonics. Figures in second column refer to the frequency band eliminated by the filter setting.

[1] Fletcher, Harvey, Physical Criterion for Determining the Pitch of a Musical Tone, *The Physical Review*, March, 1924.

It is seen that even when the fundamental and all the harmonics up to and including the seventh were eliminated, the pitch remained the same although there was a very marked change in quality. It was not until the sound had completely lost its character and become merely a noise, after the elimination of all the components up to and including the ninth harmonic, that the pitch changed to the extent of being characterized as "uncertain."

Similar eliminations of the upper harmonics, leaving the lower components intact, produced no change in pitch but varying changes in quality, as was to have been expected. In the last observation recorded in this table, the eliminations were made from both ends of the spectrum, leaving only the narrow intermediate band of frequencies of from 500 to 1,000. Again, although the fundamental, first and second harmonics and the entire range above the sixth harmonic were missing, no change in pitch occurred, but there was a marked change in quality.

Observations on other vocal sounds and also on the sounds of various musical instruments were made with generally similar results. Fletcher summarizes some of his results as follows:

"In general, neither the quality nor the pitch of notes from a rich baritone or contralto voice is appreciably affected by eliminating the fundamental and the first two or three overtones. If, however, higher overtones are eliminated, the musical quality (in particular the richness) is notably affected, and this is true even though the omitted overtones are all above the fifteenth. The high harmonics do not seem to be so essential for good quality in a soprano voice. Experimental tests showed the rather unexpected result that the elimination of all the harmonic frequencies above 2,000 cycles affects the musical quality of a bass, a baritone or a contralto voice to a greater extent than the quality of a high soprano voice.

"The . . . quality of the principal musical instruments is much more seriously affected by the elimination of the lower parts of their characteristic sound spectra than the quality of the sung vowels by a similar elimination. In any case, such eliminations do not change the pitch for this remains constant as long as the filtered sound can be recognized as a musical tone."

In another striking set of experiments the sounds were built up synthetically by combining in various ways, in a telephone receiver, currents of ten different alternating current sources.

These gave, accurately, frequencies of 100, 200, 300, etc., up to 1,000 cycles per second. The currents were so adjusted that the pressure amplitudes of all the components of the sound emitted were equal, and the arrangement was such that any of the generators could be switched out of the circuit and its component thus completely eliminated. Here no question of any trace of the original components being left in the sound with a possible suggestive influence on the ear could arise, for the components were definitely in or out according to the position of the switches.

When all of the generators were impressed on the circuit the resulting full tone, as was to be expected, had a definite pitch corresponding to 100 cycles. But the elimination of the 100-cycle component produced no noticeable effect either in pitch, loudness or quality. The elimination of other single components had no effect on the pitch and almost none on the quality. It is particularly significant that with all of the components adjusted to the same pressure amplitude the withdrawal of the fundamental had the least effect of all.

Even when the first seven components were eliminated, leaving only those having frequencies of 800, 900 and 1,000, the pitch remained definitely that of a 100-cycle tone. The withdrawal of the 800 frequency, leaving only the 900 and 1,000, left the remaining sound recognizable as two separate tones, of pitch corresponding to those two frequencies, respectively; but even here the fundamental subjective tone of 100 cycles was still plainly discernible though weak.

In general it was found that with the fundamental eliminated any three consecutive components, such as 600, 700 and 800, were sufficient to give a complex tone corresponding to a frequency of 100 cycles. When four consecutive components were sounded, the fundamental subjective tone was very prominent. When all were sounded, this fundamental seemed to be louder than any of the other components and completely dominated the tone.

Again, it was observed that if frequencies of 200, 400, 600, 800 and 1,000 were used, the pitch was that corresponding to 200 cycles, and the same was true if the 200- and 400-cycle components were eliminated. Any two consecutive pairs, differing by 200, gave faintly the sujective tone corresponding to 200 cycles.

Up to this point in these experiments the subjective tone supplied by the ear has always been that corresponding to the

common difference of the series of frequencies. It was found, however, that when frequencies of 100, 300, 500, etc., were combined, the resulting sound was a noise of no definite pitch. The same was true of other combinations like 100, 400, 700 and 1,000, and also 100, 500 and 900. These latter and other experiments seemed to show that the common difference of the arithmetical series of frequencies determined the pitch of the sound only when the components in the series were exact multiples of that common difference.

We are thus forced to the conclusion, contrary to previous belief, that the fundamental vibrations of a sound do not always determine the pitch. The experiments seem to indicate that the real factor in determining the pitch is the common difference in the arithmetical series of frequencies, when each of the component frequencies is an exact multiple of this common difference.

Fletcher accounts for the observed facts by the non-linear response of the ear to forces applied at the eardrum—that is, the non-linear transmission characteristics of the middle and inner ear. All of the summation and difference frequencies and the harmonic frequencies produce nerve stimulation resulting in the *subjective tones*. He illustrates by the diagram of Fig. 67 in which chart *a* shows the spectrum of the sound produced by the ten sources of alternating current used in making the series of experiments just referred to. It will be remembered that all the components were adjusted to have the same pressure amplitude value as applied to the ear drum. Chart *b* shows what the spectrum at the auditory nerve terminals would be when all the components were sounding together, if the ear had a linear response. This is because the higher frequencies carrying more energy produce very much greater sensation than the lower frequencies at the same pressure amplitudes. But as the ear has a non-linear response, the actual spectrum at the auditory nerve terminals was quite different and is shown in chart *c*. Here the fundamental of 100 cycles is greatly reinforced by the nine successive difference tones of that frequency, corresponding to the nine 100-cycle intervals between the successive components above it. Likewise, the first harmonic component of 200 cycles is greatly reinforced by the four 200-cycle intervals between the 200, 400, 600, 800 and 1,000 components above. It is seen that these subjective tones added to the fundamental are of far greater importance than the fundamental frequency actually

present, and it was no doubt these added subjective components that actually dominated the sound. This is indicated in chart *d*, which is the actual spectrum of the sound sensed by the ear when the fundamental and succeeding three frequencies were actually eliminated from the sound impressed on the eardrum.

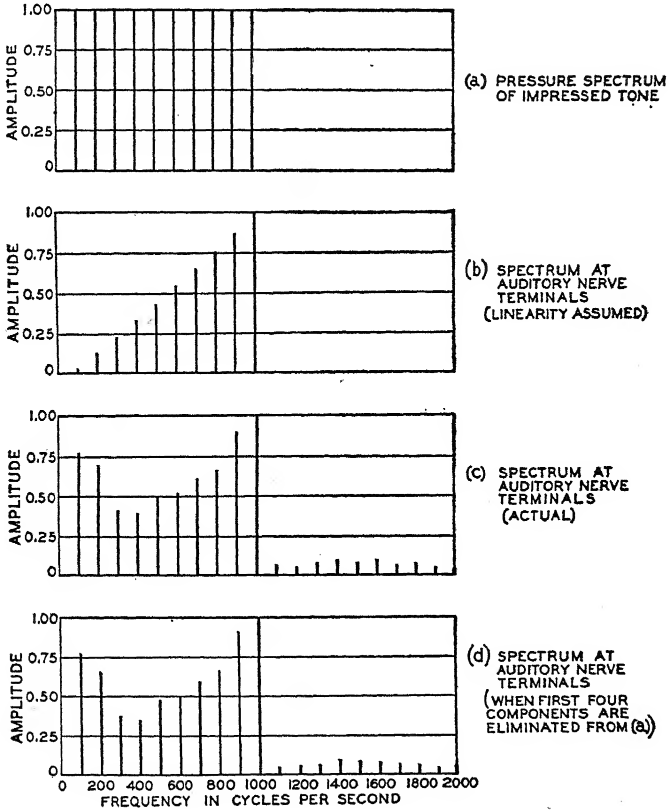


FIG. 67.—Comparison of impressed and sensation spectra. (*Courtesy of Bell Telephone Laboratories.*)

This ability of the ear to supply, subjectively, fundamentals and other components at the lower end of the spectrum, which do not exist in the impressed sound, helps to explain why the intelligibility of speech suffers less from the elimination of bands of frequencies at the lower end of the spectrum than from the upper. The ear can better supply, subjectively, omissions at

the lower end than at the upper. Again it throws some light on the puzzling phenomenon of the recognizable though distorted transmission of music which is known to contain certain essential tones in the register well below middle *c* (256) by a radio system of such poor design as to make it impossible for it to reproduce any frequencies below 300. In such cases the ear supplies, subjectively, some of the tones which the radio system, due to its poor design, fails to transmit.

It is found that the ability of the ear to perceive a sound is often lessened or destroyed by the presence of another sound. A tone of one frequency may "mask" or obscure a weaker tone of another frequency, the weaker tone not being heard at all until it has been increased to an intensity greater than that at which it would be just audible if sounding alone. The difference between the minimum or threshold intensity of a tone sounding alone and its minimum perceptible intensity in the presence of another or "masking" tone may be little or great, depending principally on the relative frequencies and relative intensities of the two tones.

The amount by which the threshold value of a tone sounding alone has to be increased to make it just perceptible in the presence of another tone is called the "threshold shift." It may be expressed in the number of loudness units between the two thresholds.

In general it may be stated that the masking effect of one tone by another is greater for tones near together in frequency than for those whose frequencies are widely separated. An exception to this statement must be made, however, in the case of two tones so nearly alike as to cause beats. Within the range of beats the masking effect is lessened, for here the presence of the beats helps to make the weaker tone perceptible. In general, also, a loud masking tone will obscure tones of higher frequencies than its own very much more effectively than those of lower frequency. For a weak masking tone, however, the masking is about equally effective on frequencies above and below its own.

Figure 68 is drawn from one of many similar charts by Fletcher[1] and is based on experimental research carried out principally

[1] FLETCHER, HARVEY, Physical Measurements of Audition and Their Bearing on the Theory of Hearing, *Journal of the Franklin Institute*, September, 1923.

by Messrs. Wegel and Lane[1] in the Bell Telephone Laboratories. It will serve as an example to illustrate the general statements just made concerning masking. In this the different curves show the masking of tones having frequencies ranging from 400 to 4,000 cycles, by masking tones having a fixed frequency of 1,200 cycles but of five different degrees of loudness. The figure opposite each curve shows the intensity of the masking

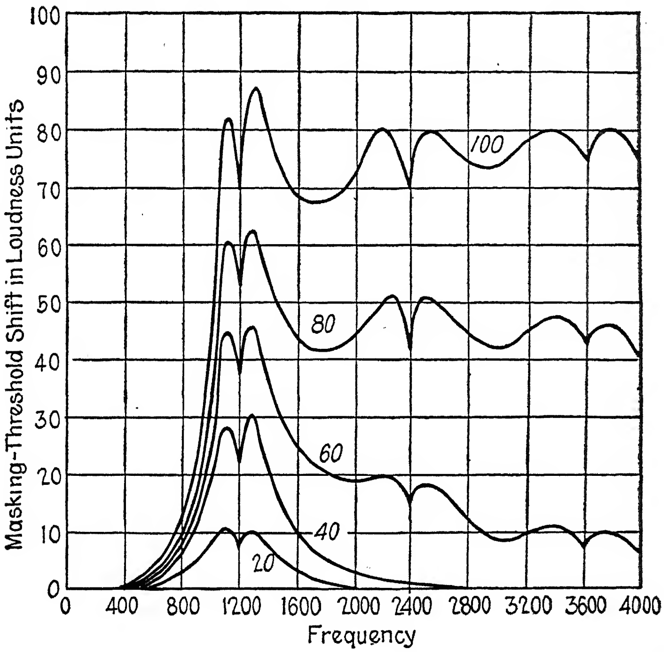


Fig. 68.—Masking of tones of various frequencies by 1,200-cycle tones of different loudness.

tone expressed in loudness units above its threshold. The abscissas show the frequencies of the masked tones in cycles per second, and the ordinates the masking effect upon them. The masking effect is expressed in the number of loudness units the masked tone must be raised above its ordinary threshold value to render it just audible in the presence of the other sound.

To illustrate the interpretation of these curves: The presence of a 1,200-cycle masking tone 60 units loud will make it necessary

[1] Wegel, R. L. and C. E. Lane, Auditory Masking and Dynamics of the Inner Ear, *The Physical Review*, February, 1924.

to raise the threshold value of a 1,600-cycle tone about 25 units before that tone becomes audible. If the masking tone is 100 units loud, the required threshold shift of the 1,600-cycle tone will be 68 units, this increase in its intensity being required to make it barely audible in the presence of the masking tone.

In all cases it is seen that the greatest masking occurs when the frequencies of the masked tones are close to that of the masking, yet not close enough to produce beats, and almost equally so for those just below and just above it. For the weak masking tone 20 units loud, the curve slopes off almost symmetrically on each side, the masking rapidly diminishing to zero as the frequencies fall below or rise above that of the masking tone. As the masking tone becomes louder an increasing difference is to be noted as between the effect on tones below it and on tones above it. For the masking tone 80 units loud, for instance, there is a marked masking effect on all tones of frequencies above the 1,200-cycle tone, while on the other side the masking rapidly drops to zero as the frequencies of the masked tones decrease. In the case of the very loud (100-unit) masking tone there is little diminution of the masking throughout the entire range above the masking frequency, while, as before, the effect rapidly disappears for tones of frequencies below it.

The marked depression in each of the curves of Fig. 68 for frequencies close to that of the masking tone show the decrease of masking that always occurs when the tones are close enough together to cause beats.

We may now inquire briefly into the cause of masking in the light of the dynamic theory of the action of the inner ear, considered earlier in this chapter. That theory assumes that the entire basilar membrane vibrates for every tone, that only vibrations beyond a certain amplitude cause nerve stimulation and that the brain interprets the sensation by the portion of the membrane where the nerves are stimulated. Since the amount of stimulus of any nerve is supposed to depend upon the degree of motion of the basilar membrane at the point where the nerve terminates, it seems reasonable to assume that when the membrane is already vibrating under the influence of one tone, no additional stimulus will be caused by a second tone unless the movements caused by it extend beyond the limits of the movements caused by the first.

Figure 69 is adapted from a hypothetical diagram used by Wegel and Lane in their discussion of the amplitude of vibration along the basilar membrane. On this the abscissæ represent distances along that membrane measured from the oval window and ordinates the amplitude of vibration of the various points on the membrane. The height of the horizontal dotted line above the horizontal axis represents the minimum audible amplitude, which is assumed to be constant for all points along

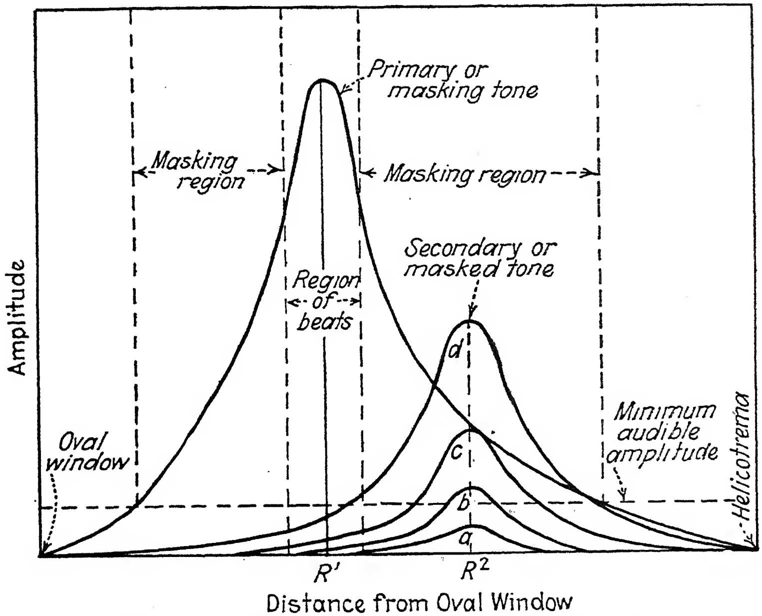


FIG. 69.—Hypothetical curve of vibration of basilar membrane.

the membrane. Vibrations below this threshold line produce no sensation and, therefore, of course, no masking.

The major curve of Fig. 69 may be taken as indicating the amplitude of vibration throughout the length of the basilar membrane caused by a primary tone which we may consider the masking tone. The two vertical lines on each side of the point of maximum amplitude and close to it are arbitrarily chosen to mark the limits of the region within which two different tones will be so close together in pitch as to cause beats. Within this region masking will be lessened because of the beats, as already mentioned. Obviously, also, the two outer vertical dotted lines on this diagram will represent the most remote

points on the basilar membrane between which masking by this tone can occur, for values of the masking tone amplitude outside of these limits lie below the line of minimum audible amplitude. Similar curves, *a*, *b*, *c* and *d*, show the vibrations of the basilar membrane produced by a weaker secondary tone of another pitch, these four curves representing four different intensities of the same tone.

The weakest of the values shown on this secondary tone, curve *a*, would, obviously, be inaudible even in the absence of the masking tone, because its maximum amplitude lies below the limit of audibility. The second intensity of the secondary tone, curve *b*, would be audible if it alone were causing the basilar membrane to vibrate, for its maximum amplitude rises just above the normal threshold line. In the presence of the primary tone, however, it will not be audible, because the corresponding part of the basilar membrane is already being vibrated with greater amplitude by the primary tone, and, therefore, no added stimulation of the nerve centers at that point is produced.

We know, however, that if the amplitude of the secondary tone is gradually increased, a point will be reached where it will become audible in spite of the masking of the primary tone. It is assumed that this will occur when the maximum amplitude, due to the secondary tone, is about equal to the amplitude of the vibration already produced at the corresponding point by the primary tone. Curve *c*, therefore, would represent a secondary tone whose amplitude was just at the threshold of audibility in the presence of the masking tone. For any intensity above this, such as that represented by curve *d*, the tone will be perceptible in the presence of the masking tone, but its threshold will have been shifted, as indicated on the diagram.

The phenomenon of masking, once a stumbling block in the path leading to an understanding of the theory of hearing, now proves to be useful in gathering information helpful to such an understanding. Owing to the inaccessibility of the basilar membrane in living persons, it is, of course, necessary to resort to indirect means in observing its vibrations and the sensations caused thereby. The chart of Fig. 69 may be referred to in discussing a method that has been developed for the exploration of its vibratory movements at differing points along its length caused by sounds impinging on the eardrum.

The apparatus used in making such explorations consists essentially of two generators of sinusoidal alternating current, each delivering its current to a common telephone receiver and producing therein two pure tones of the desired frequency. Auxiliary apparatus in connection with each generator circuit permits each of the tones to be brought to any desired intensity and measured.

We will say that one of these generators is producing a primary tone of known intensity, and that it is desired to determine the amplitude of vibration produced by it at various points along the length of the basilar membrane. In other words, the form of such a curve as the major curve of Fig. 69 is the matter for determination.

The point of maximum amplitude on such a curve is at once obtainable. The known intensity gives its ordinate, and the known pitch, in connection with such information as is given in Fig. 62, shows the point at which maximum vibration will occur. The only other directly obtainable points are at the two ends of the basilar membrane, when the ordinates, of course, are zero. For each other point of the unknown curve a secondary or exploring tone is used. The pitch of this exploring tone for any single determination will correspond to a known point on the basilar membrane (Fig. 62). Its intensity at the time when it begins to be noticeable above the sound of the primary tone will give the desired amplitude of vibration of that point that is being caused by the primary tone. This is true because it is at that intensity that the amplitude of the exploring tone begins to rise above the amplitude caused by the primary tone.

Thus referring again to Fig. 69, the known frequency of the tone under investigation locates the point R_1 along the basilar membrane where the maximum vibration is occurring. The maximum amplitude at that point is determined in appropriate units by direct measurement of the current causing the tone. For any other point, such as R_2, an exploring tone (curves a, b, c, d) is applied at proper frequency to correspond to that point. This is slowly increased in intensity until it reaches a value where it becomes just audible above the masking of the primary tone. At this point its maximum amplitude will, in accordance with the theory advanced, be a fair approximation of the amplitude at which that point on the basilar membrane is already being vibrated by the tone under investigation. In this way, using exploring tones of other frequencies, the amplitude of

vibration of other points on the basilar membrane may also be determined.

Figure 70 shows a plotting, by the method just described, of the amplitudes of vibration along the basilar membrane caused by a loud pure tone of 400 cycles. Nearly half the length of the membrane was vibrating with sufficient amplitude to be perceptible and measurable. The point of maximum vibration lies about 21½ millimeters from the oval window, corresponding to the position of maximum response for that frequency as shown in the diagram of Fig. 62.

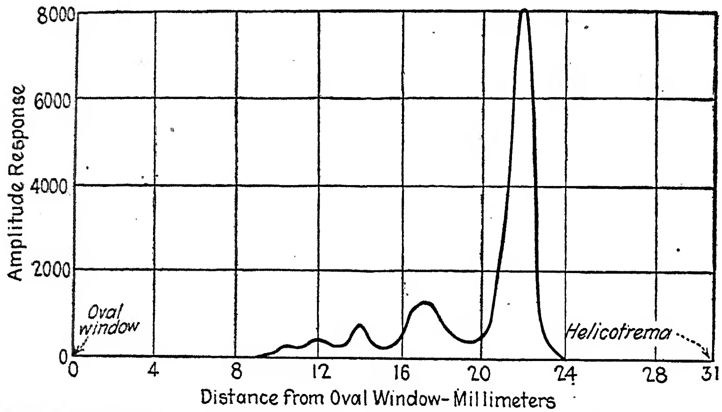


FIG. 70.—Response of basilar membrane to loud tone of 400 cycles (loudness of 78).

The curve also shows several other maxima, and further comparison with Fig. 62 will show that these correspond to frequencies of 800, 1,200, 1,600 and 2,000 cycles, the first, second, third and fourth harmonics of the impressed *pure* tone.

It is characteristic of the mechanism of the ear thus to introduce subjective harmonics of loud pure tones into the vibrations of the basilar membrane and thus into corresponding sensations in the brain. These harmonics are sensed even though careful investigation shows that they are not appreciably present in the sound vibrations reaching the ear through the air. Wegel and Lane are inclined to ascribe this non-linearity in the response of the basilar membrane, which produces these subjective harmonics, to the mechanism of the middle ear, possibly to friction in the point between the hammer and anvil when violently agitated.

Whatever their cause, the presence of these subjective harmonics goes far to explain why a loud tone masks tones of higher frequency than its own more effectively than it masks those of lower frequency. As is clear from Fig. 70, the portions of the basilar membrane thrown into perceptible vibration by the loud pure tone extend far into the region of higher frequencies and only a slight distance into the region of lower frequencies. Within these regions the nerves necessary to perceive a second tone are already stimulated, and the second tone cannot be heard until its intensity rises above the new threshold thus created for it.

Although these points of maximum amplitude on the basilar membrane do not correspond to any vibrations actually present in the impressed tone, the fact that they are points at which the membrane is actually vibrating with maximum amplitude for their respective regions suggests that they may exhibit exactly the same masking characteristics on other tones as the maximum point corresponding to the impressed tone. This is manifest from a further consideration of Fig. 68. For the lower intensities of the 1,200-cycle masking tone, corresponding to 20 and 40 loudness units, no subjective harmonics appeared, the curves showing a smooth descent for the higher frequencies. For the higher intensities of 60, 80 and 100 units, however, there are distinct depressions in the masking values at frequencies of 2,400 and 3,600 cycles, the first and second harmonics of the impressed tone. This showed that the ear mechanism had introduced these harmonics and that a lowering of the masking occurred in the regions when the frequencies of the masked tones were close enough to the harmonic frequency to cause beats.

The regions on each side of these depressions marking the harmonic points, and close to them, show maximum masking in each case, as did the ranges of frequency above and below the frequency of the impressed tone.

For the very loud (100-unit) masking tone of Fig. 68, the masking in the region of the harmonics was almost as great as that of the 1,200-cycle tone. In fact, for this loudest intensity of the masking tone, the curve shows that it would be necessary to raise a masked tone of any higher frequency up to 4,000 cycles, from about 70 to 80 units above its normal threshold value, before it could be heard at all.

In concluding this chapter on the sense of hearing the time-honored theory that the ear senses differences in the quality of sounds by detecting differences in the form of the impressed sound wave may be briefly alluded to. The ear simply does not work in that way. A complex sound wave may have its components shifted about in phase to any degree without producing the slightest noticeable difference in the quality of the sound. Yet each shift of phase may produce a change in the resultant wave form. The ear senses quality by the number and relative amplitudes of the components of the complex wave, paying no attention, so far as quality is concerned, to the phase relationship of the components.

The ear—or more properly the two ears—do take cognizance of phase differences in exercising the "binaural sense." This is the sense by which we are made aware, in a very general way, of the direction from which a sound comes. Owing to their slightly different positions relative to any source, sounds may reach the two ears in either the same or in different phases, and it is by some sensing of this difference in phase by the two ears that the binaural sense functions. One interesting investigation of binaural sensitivity was made by Mr. A. L. Bennett[1] at Union College. He used a rapidly rotating contact adapted to engage successively two adjustable arms so arranged that he could produce pulses in two head receivers at a measurable time interval apart. He found that anyone with both ears normal could tell, for intervals greater than about one one-thousandth of a second, that the two pulses were not simultaneous and could tell which ear received its pulse first. For shorter intervals the pulses seemed to be simultaneous—that is, there was no sensation of separate pulses. The observer got the sensation of a source of sound in or just above the top of the head, and for simultaneous pulses this source seemed to be exactly in the center. If the right ear was caused to receive the pulse slightly in advance of the left, the sensation conveyed to the observer was an apparent shifting of the source slightly to the right. An advance of the pulse in the left ear was sensed as a shifting to the left. Ability to detect such a shift varied in different observers, the best detecting a shift of a few millionths, the worst only slightly less than a thousandth-second interval.

[1] BENNETT, A. L., Efficiency of the Ears as a Means of Detecting Short Time Intervals, *Journal of the Optical Society of America*, vol. 14, p. 342, 1927.

The average was of the order of a ten-thousandth or less. With average keeness of binaural sense, assuming the ears to be half a foot apart and the velocity of sound to be 1,100 feet per second, the average person should be able to locate the direction of a source of sound to within less than 15 degrees, provided there are no reflections or other confusing phenomena involved.

A person who has hearing in only one ear has no binaural sense, just as a person with sight in only one eye has no stereoscopic sense. The binaural sense of hearing, like the stereoscopic sense of sight, is made possible by hearing (or seeing) the same thing simultaneously from two slightly different points. The binaural sense is not very well understood and has not been generally useful in commercial telephony in which the sound transmission is essentially of a monaural character. Perhaps, however, as more is known about it, it will appear of greater importance in the electrical sound transmission arts. One of the defects of talking moving pictures, for instance, is the sense of unreality, noticed by some, in that the sound seems to come from a fixed spot while the characters are moving about.

CHAPTER VII

THE KINDS OF SOUND

The general characteristics of sound vibrations have been discussed and also the way in which these vibrations are received by the ear and converted into sensations to be interpreted by the brain. In this chapter sound vibrations will be classified and discussed with particular reference to the considerations involved in their conversion into electrical waves, their transmission as such to distant points and their final reconversion there into sound waves.

There is such an infinite variety of sounds merging into each other by such indefinable shadings that it is difficult to classify them even broadly. We may at least roughly divide them into two classes, however, noises and musical tones, or merely noises and tones. Even for this simple classification it is difficult to draw a sharp dividing line. To one person a sound may be a tone to another a noise. Again, a sound may be called a tone at one time, and, at another time, under other circumstances, the same person may characterize it as a noise. These differences may be due to various causes.

To illustrate: The sounds of Wagnerian opera are mere noise to some but music to others. The difference is one of taste or musical education. Again, the sound caused by dropping a single stick of wood on a floor would probably be called a noise, but if a number of the same sort of sticks of prearranged sizes are dropped successively in proper order, the sounds are easily recognized as tones. In fact, tunes may be played in this way. Here it is the relationship among the successive sounds, the context, that causes the person to characterize the same sound as a noise in one case and a tone in the other. The difference is in the circumstances, not in the sound.

Sometimes the attempt is made to distinguish between noise and tone according to whether the sound is disagreeable or pleasing. This is clearly unsafe ground. Noise is not necessarily unpleasant. The popping of firecrackers causes acute delight to the small boy, if not to his anxious mother.

A distinction between noise and tone which probably comes nearer the facts is that noise is caused by aperiodic and tone by periodic vibrations. Judging from the resultant wave forms of continuing noises and tones, this is a good distinction, for, as will be shown, the wave form of a noise apparently lacks the regularly recurring properties which characterize the tone. But we must go deeper than mere periodicity or the lack of it for the real difference between noise and tone. We know that we can produce unmistakable noises by a heterogeneous mixture of unrelated and rapidly changing periodic vibrations.

A more satisfactory distinction, which physicists are adopting, takes into account the capabilities of the human ear as well as the physical make-up of the sound. Under this view a noise is a sound which is either so short in duration or so complex in wave form that the ear cannot understand it and, therefore, cannot recognize in it any definite pitch. Tones, on the other hand, are sounds having such continuity and such regularity of wave form that their characteristics may be appreciated by the ear.

Under this conception the differentiation becomes one of understanding, and this accounts, at least in part, for the fact that different persons will give different interpretations to the same sound, or that the same person will change his interpretation after gaining experience or after hearing the sound amid a context of other sounds.

Noises, on the other hand, represent the unorganized and disorderly class of sounds. Indeed, the very derivation of the word "noise" seems to suggest confusion. Tones, on the other hand, are the organized and orderly sounds which the ear can interpret and understand. Of these speech and music are principally composed, although, as will be shown, it cannot be said that noise is completely absent from either.

Telephony is concerned primarily with speech but also with music and noise. It is concerned with speech and music because it is desired to transmit them; with noise because it is desired to suppress it. How most exactly to transmit those sounds which it is desired to transmit and to suppress those which it is desired not to transmit is the fundamental problem of either wire or wireless telephony.

Speech.—Of the three classes of sound, speech, music and noise, speech is of outstanding importance not only in the arts concerned in the electrical transmission of sound but also in

the ordinary affairs of daily life. Perhaps it is the power of speech which most strikingly distinguishes man from all other living creatures.

Speech is built up of articulate sounds so combined and merged from one to another as to form words, phrases and sentences capable of conveying meaning and of being understood. It is composed principally of tones variously shaded and intricately mingled. Some of the elementary sounds of speech, if occurring

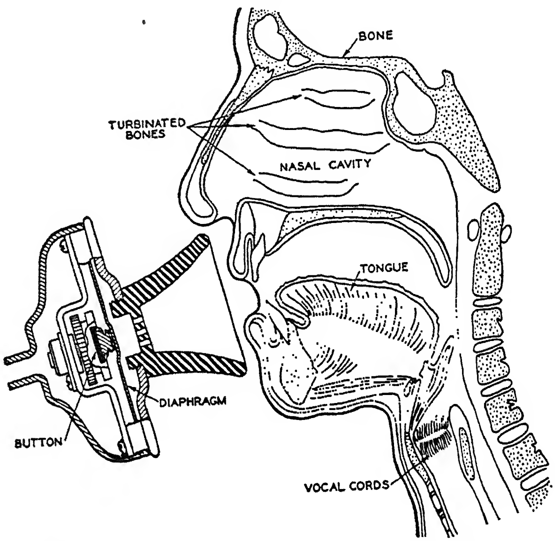


FIG. 71.—Cross-section of vocal organs and of telephone transmitter. (*Courtesy of Bell Telephone Laboratories.*)

separately, would undoubtedly be classed as noises, but, through long practice, the ear has become sufficiently educated to understand and interpret them.

The only natural source of all speech is, of course, the human vocal organs. A cross-section of these and of a telephone transmitter in proper position to receive their sounds is shown in Fig. 71.

The primary tones of most, but not all, of the speech sounds are caused by vibration of the vocal cords, a pair of muscles stretched across the larynx at the base of the throat in such manner as to form a straight slit between them. These are caused to vibrate by air currents forced in and out through this slit by the bellows action of the lungs, in much the same manner as a rubber band stretched across the outlet of a toy

balloon is made to vibrate by the out rush of air from the balloon. Instead of being confined practically to the production of a monotone, however, as is the balloon whistle, the vocal cords, without the other modifying influences to be mentioned, are capable, by their own muscular action, of varying their own tension and form. They thus have, within themselves, the capability of producing some variations in the pitch and quality of the sounds they emit.

The train of relatively simple sound waves started by the vibrations of the vocal cords then traverse the passages of the throat, mouth and nose, which act as resonant chambers, the size and shape of which are subject to wide variation due principally to the movements of the tongue and lips.

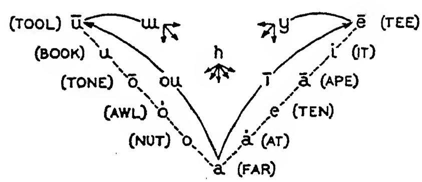
These latter modulations caused by variations in the size and shape of the vocal chambers are, as will be shown, of very much more importance in the production of speech sounds than are those modulations due to changes in the tension and form of the vocal cords themselves. In fact, it has been proved recently that a man with no vocal cords at all may learn to speak quite understandably if a whistle, similar to that of a toy balloon, is provided in such manner as to be actuated by the air currents of the lungs and its sound modulated in the usual way by variation in the cavities of the mouth and nose. Again, we are all familiar with the fact that we can communicate with each other in whispers, in which case the vocal cords are not brought into play at all.

These general facts in connection with a large amount of experimental research data that will be referred to later lead to the following conclusions: That the vocal cords produce the lower range of frequencies, that is, those below about 500 cycles; that these lower frequencies may be considered as the foundation or carrier waves of speech sound; that they carry the principal energy of speech but are of comparatively little importance in contributing to its intelligibility. On the other hand, the finer modulations and higher frequencies produced in the oral and nasal cavities, largely by movements of the tongue and lips, contribute little energy to the sound but much to those characteristics which make speech intelligible.

Variations in loudness are caused principally by changes in the force of the air streams supplied by the lungs but also by bringing the resonant chambers more or less into play. In

addition, the action of the tongue and lips serves to control the starting and stopping, thus breaking up the wave train into syllables and words. They also serve, without the aid of the vocal cords, in producing certain of the speech sounds. Even the teeth are brought into play in making the very high-frequency hissing sounds that are important elements of speech. The lungs, the vocal cords and the vocal cavities together with the tongue, lips and teeth combine to make by far the most versatile

1. VOWELS



2. COMBINATIONAL AND TRANSITIONAL VOWELS

ω-y-ou-ī-h

3. SEMI VOWELS

l-r

4. STOP CONSONANTS

VOICED	UNVOICED	NASALIZED	FORMATION OF STOP
b	p	m	LIP AGAINST LIP
d	t	n	TONGUE AGAINST TEETH
j	ch	-	TONGUE AGAINST HARD PALATE
g	k	ng	TONGUE AGAINST SOFT PALATE

5. FRICATIVE CONSONANTS

VOICED	UNVOICED	FORMATION OF AIR OUTLET
v	f	LIP TO TEETH
z	s	TEETH TO TEETH
th (THEN)	th (THIN)	TONGUE TO TEETH
zh (AZURE)	sh	TONGUE TO HARD PALATE

Fig. 72.—Classification of speech sounds. (*Courtesy of Bell Telephone Laboratories.*)

and flexible of all sound producing instruments. Nothing approaching it in these respects is to be found in nature or in the works of man.

The nature of language and the classification of its sounds are interestingly dealt with in an article[1] by Mr. R. L. Jones of the Bell Telephone Laboratories. He groups the thirty-six letter sounds of English speech into five classes: (1) pure vowels; (2) combinational and transitional vowels; (3) semivowels; (4) stop consonants; and (5) fricative consonants. This classification he analyzes further in Fig. 72.

[1] Jones, R. L., The Nature of Language, *Journal of the American Institute of Electrical Engineers*, p. 321, 1923.

The triangular diagram at the top of this figure shows the various sounds of the pure vowels, *a*, *e*, *i*, *o* and *u* and also the various combinations in which these sounds are used in connection with the transitional vowels, *w*, *y*, *ou*, *ī* and *h*. The vocal cords are always brought into play in speaking the vowel sounds. The low fundamental and overtones of the sound are then produced, but the characteristics necessary to determine which vowel is being spoken are furnished by the resonance of the vocal chambers as controlled by the tongue and lips already referred to. Thus for the same action of the vocal cords we may pass through the entire range of vowel sounds by merely varying the positions of the tongue and lips.

This procedure is about as follows: The vowel *u* (as in tool) at the upper left corner is formed by rounding the lips, drawing back the tip of the tongue and raising it at the back in such a way that the throat is almost closed off and the mouth forms a single large resonant cavity. As we come down the left-hand side of the triangle pronouncing the sounds indicated, the lip opening widens and the jaw is lowered. We are still depending on single resonance until the bottom of the triangle is reached. In pronouncing the sound *a* (as in far), the lips are wide open, the jaw is dropped and the tongue is only slightly raised at the back. This gives a strong reenforcement in the neighborhood of 1,000 cycles. With the *a* (as in at), the lips are still wide open, the tongue lies flat in the mouth and the mouth and throat form connected cavities of nearly equal size. Here there is double reenforcement due to the two cavities, these being at about 800 and 200 cycles. As the vowels on the right-hand side of the triangle are pronounced, proceeding upward, the separation of the lips becomes smaller and the tongue is raised first in the center and then farther forward. In all of these sounds on this side of the triangle, double resonance is found due to the cavities in front of and behind the tongue. When the upper right corner of the triangle is reached, the lips sounding *e* (as in tee) are drawn to form a wide slit and the tongue is raised in front until its ridge is close to the roof of the front of the mouth. Double resonance here is caused by the large cavity formed by the back of the mouth and the throat and by the small space over the tongue at the front of the mouth. Here the two frequency reenforcements are in the vicinities of 300 and 2,500 cycles.

Some of the vowel sounds are called "transitional," because they start with one vowel sound and end with another. Thus, in pronouncing $\bar{\imath}$ the mouth is formed as if to say *a* (as in far) and then rapidly altered to sound $\bar{e}$ (as in tee). This is indicated on the diagram by placing $\bar{\imath}$ in the middle of the long curved line at the right of the triangle extending from *a* to $\bar{e}$. Similarly, the sound of *w* is made by forming the mouth to say $\bar{u}$ and then suddenly changing to any of the pure vowels. The sound of the transitional letter *h* may be prefixed to the various vowel sounds, as indicated by the arrows, by separating the vocal cords initially and beginning the sound by forcible expulsion of the breath.

The sounds of *l* and *r* are thrown into the third class of semi-vowels because they partake of some of the characteristics of the vowel sounds.

The fourth class, "stop consonants," is produced in connection with a stoppage in some part of the mouth. These are "voiced" in case they require the use of the vocal cords, "unvoiced" if they are made without using the vocal cords and "nasalized" if the nasal passage is brought into play. The column at the right of the table of stop consonants shows how the stop is formed in each case.

The fifth class of speech sounds, "fricative consonants," is characterized by the rushing sound of the breath through variously formed openings of the mouth. Thus, for *v*, a voiced sound, the outlet is formed by placing the lips against the teeth, while *f* is produced through the same sort of an outlet but unvoiced. The voiced and unvoiced fricatives and the methods of forming the air outlets in producing them are shown at the bottom of the table in Fig. 72.

The process of uttering speech sounds is thus seen to be very complicated. The variations from moment to moment in the tension and form of the vocal cords, in the size and shape of the resonant cavities in the path of the air stream, in the force by which the air stream is supplied by the lungs, in the character of the outlet of the mouth through which the sounds emerge and in the manner of stopping performed by the tongue and lips cause an infinite variety in the moment-to-moment changes in the loudness, pitch and quality of the resulting wave trains in the air. An oscillographic record of the form of such a wave train for a single word—"farmers"—is shown in Fig. 73. At

the bottom of this figure is shown, to the same scale, the record of a simple tone of 500-cycle frequency.

The abscissæ of this record represent the passage of time and the ordinates variations in pressure amplitude. By using the horizontal distances on the record of the 500-cycle tone as a standard, the frequencies of the various parts of the record of the spoken word may be determined with fair accuracy.

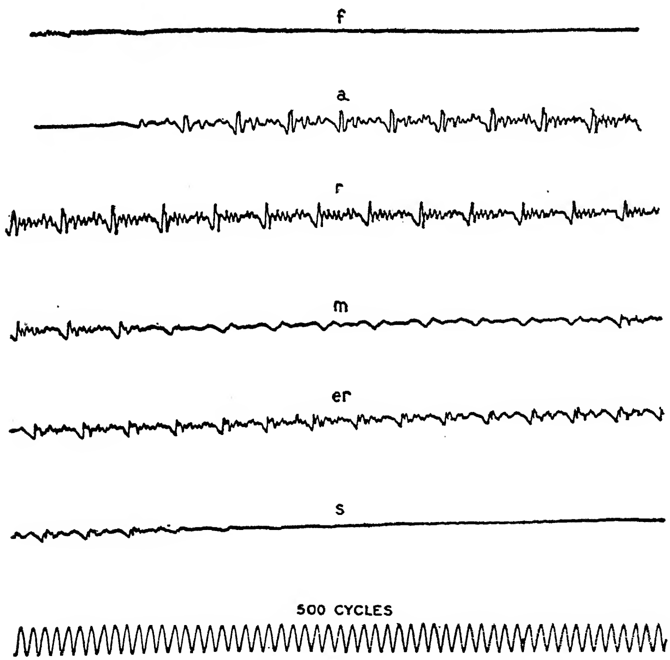


FIG. 73.—Wave form of the word "farmers." (*Courtesy of Bell Telephone Laboratories.*)

Here some of the characteristics already alluded to are shown graphically. The fricative consonants *f* and *s* are each seen to produce waves of very low amplitude and high frequency. Their energy is low and their pitch high. The succeeding sounds of the vowel *a*, the semivowel *r* and of the stop consonant *m* are each characterized by fundamental tones of pitch corresponding to about 120 cycles, all having much greater pressure amplitudes than those of the fricative consonants with which the word began and ended. Upon reaching the semivowel *er* the voice was slightly raised in pitch, the frequency of the fundamental rising to about 130 cycles.

The record shows for each of the sounds of this word, except the *f* and *s* sounds, a number of harmonics superimposed on the fundamentals, in each case producing a wave form of considerable complexity. The reason the higher harmonics were not also apparent on the records of the two fricative consonants was simply that their energy was so small and their frequency so high that the oscillograph could not follow and trace them. In fact, its capabilities were taxed in following the fundamentals of these two sounds. For exactly the same reason the more delicate ripples on the waves of the other sounds in this word failed to be recorded. They were beyond the range of sensitivity of the recording instrument.

This brings out a principal fault in the oscillographic method of recording the vibrations of speech sounds. It is of great value in showing a continuous record from one end to the other of a varying sound, but for more complete analysis, which will take into account the very delicate high-frequency low-amplitude ripples that are superposed on the grosser waves, other methods must be resorted to.

One of these other methods we may call the "selective filter method" of voice analysis. While it is not so striking, graphically, and while it is not able to follow the transitions as a speech sound progresses, it is capable of much more complete analysis of any sound of long enough duration to permit its application. It gives, so to speak, a complete cross-section of a complex sound at any one instant.

In following this method, the sound under analysis is made to produce an electrical current which varies exactly in accordance with the sound waves. Such a current will have, in proper proportion, all of the frequency components of the sound wave. An electrical filter, accurately tuned to one particular frequency, is then placed in the path of this current and in this way the presence or absence of a harmonic of that particular frequency is noted. If present, its amplitude is measured. By rapidly applying different successively tuned filters in this way, the whole range of frequency is covered and measured. The result of such an analysis is not the determination of a wave form but rather of what we may call a "sound spectrum," showing the presence and relative amplitude of each component of the sound that is found to be present.

Such a sound spectrum for the vowel *a* (as in ape) sung at a pitch corresponding to a frequency of 225 cycles is shown in Fig. 74. In this the presence of each component is indicated by a vertical line. The height of each line indicates the amplitude, and its spacing along the horizontal axis the frequency of vibration of the corresponding component. Here the occurrence of the harmonics in two distinct groups, one containing the first, second and third, and the other the sixth to eleventh inclusive, is to be noted. This double resonance, as before stated, is due to the two resonance chambers in the mouth, one in front of the tongue and the other behind it.

The spectrum of Fig. 74 is for a single vowel sound. If an indefinite number of such spectra were made to cover the entire

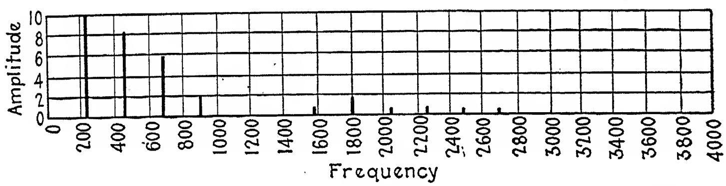


Fig. 74.—Analysis of vowel sound *a* (*ape*) sung at pitch corresponding to 225 cycles.

range of spoken sounds, and if these were combined, there would result a continuous spectrum represented by a curved line at varying heights above the horizontal axis along which frequency is measured. The height of each point of this line above the horizontal axis would represent the pressure amplitude of the corresponding frequency. Such a continuous spectrum for the sounds of the English language as they occur in conversational speech is shown in Fig. 75.

The usual range assigned to the frequencies employed in speech is from about 60 to 10,000 cycles. The spectrum of Fig. 75 is not, therefore, quite complete since it extends only to 5,000 cycles. As all the important speech frequencies lie somewhat below the 5,000-cycle mark, however, this spectrum may, for all practical purposes, be considered as covering the most useful range of speech sounds.

It is seen from Fig. 75 that the waves of greatest amplitude and those carrying the greatest amount of energy occur at the lower frequencies. This is particularly true for the frequencies between 100 and 200 cycles which, in conversation, correspond

to the fundamental tones produced by the vocal cords in uttering the vowel sounds. The high frequency end of the spectrum is occupied by the waves of low amplitude carrying little energy. They represent principally the consonant sounds. The intermediate range is occupied principally by the lower consonant vibrations and by the harmonics of the vowel sounds. Practically all of the vowel-sound harmonics lie below a frequency of 3,000, while some of the consonant sounds involve frequencies up to 6,000 and above.

An ideal system of speech transmission, whether by wire or wireless, would be one in which all frequency components were transmitted with equal efficiency, and in which no new fre-

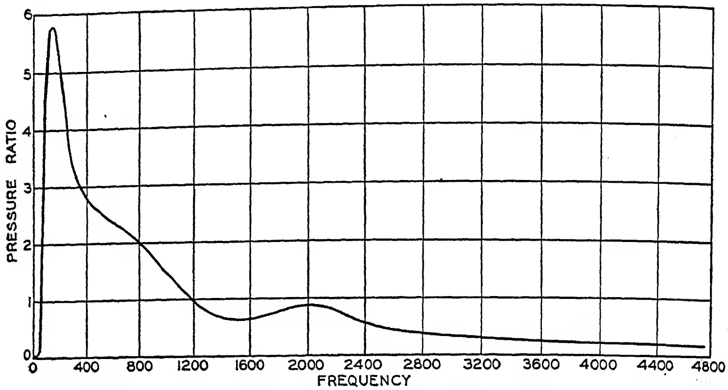


FIG. 75.—Continuous spectrum of speech. (*Courtesy of Bell Telephone Laboratories.*)

quencies were introduced. Then the reproduced speech sounds would exactly represent the originals. Here the problem of voice transmission is much more complicated than that of power transmission. The power engineer designs his system for a single frequency, whereas the telephone engineer must design his for many.

For the telephone engineer to attempt the ideal and design his system to transmit all frequencies of speech with equal effectiveness would be a vastly costly proposition—prohibitive commercially. The wider the band of frequencies he attempts to include the more costly will be the system; while, on the other hand, the narrower the band the poorer will be the quality of the speech transmission. Evidently, as in nearly all engineering problems, a compromise must be effected between allowable cost on the one hand and quality of achievement on the other.

This question of how closely the telephone system shall approach the ideal in this respect, or, in other words, how wide a band of frequencies it shall be designed to transmit, leads to another line of investigation into the characteristics of speech sounds. This inquires into the extent to which the quality of the received speech will suffer by a failure to transmit all of the frequencies involved in the original, that is, by a narrowing of the band of transmitted frequencies.

There are two general attributes of speech that are affected—*naturalness* and *intelligibility*. Perfect transmission would be both natural and intelligible. The transmitted voice would be natural in that it would sound exactly like that of the speaker, and it would be intelligible because it would be perfectly understandable. These two qualities of speech are not as closely related as might be expected. As the frequency band is narrowed naturalness suffers considerably before there is any noticeable loss of intelligibility.

Since the prime function of commercial telephony is the transmission of *intelligence*, it follows that any alterations permitted in the character of the received speech in the interests of economy should interfere as little as possible with its *intelligibility*. On the other hand, some sacrifice in naturalness does comparatively little harm, provided it does not go too far. Naturalness is, of course, desirable in that it enables one to recognize the voice of the speaker, but of itself is of distinctly secondary importance unless it departs so far from the original and becomes so unnatural as to impair intelligibility. The case is different with the transmission of music, for reasons that will be pointed out.

It is obviously of very great importance in telephony that the articulation of the speech delivered by the receiving instrument shall be good enough to render it clearly intelligible. If intelligibility fails the object of the telephone system is defeated. The real test of articulation, therefore, is intelligibility. A method has been devised for testing articulation in terms of intelligibility and expressing the results quantitatively. Paradoxical as it may seem, *unintelligible* subject matter or "material" is selected for transmission in such tests. It is found that if ordinary reading matter is used the listener unconsciously supplies from the context much that he does not actually hear, and, as a result, the test is, in some measure, a trial of the intelligence of the listener rather than of the transmission efficiency of the system.

Accordingly, the material to be transmitted consists of lists of detached monosyllabic sounds, systematically chosen to be representative of the entire range of 36 speech sounds shown in the analysis of Fig. 72. These sounds are spoken into the transmitter of the telephone system under test, and the listener at the receiving end writes them down as he hears them. The percentage of sounds correctly recorded is taken as the index of the articulation of the system.

By using the distortionless experimental telephone system, already referred to, in making these tests and introducing

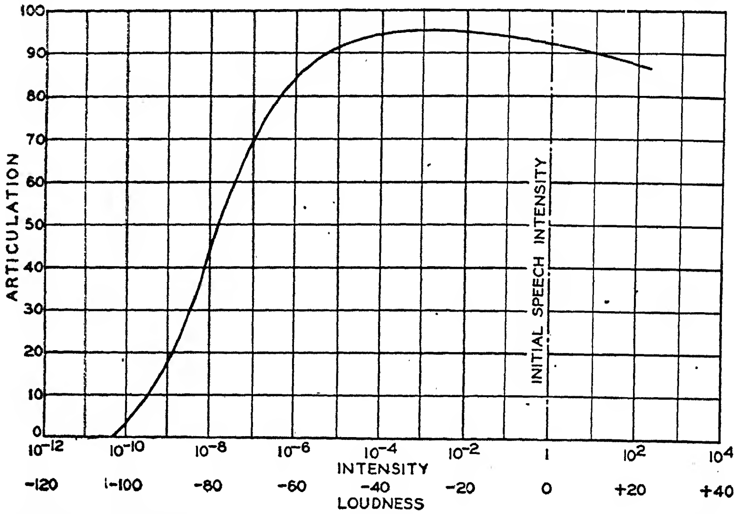


FIG. 76.—Articulation-loudness characteristic for speech. (*Courtesy of Bell Telephone Laboratories.*)

various known factors of attenuation, amplification or distortion, the resulting effects on articulation may be measured. Furthermore, this method of testing lends itself readily to the determination of which speech sounds are most liable to error in transmission and also of the relative degree in which these sounds suffer by different possible shortcomings of the system.

Figure 76 gives the results of such a test to determine how mere changes in attenuation and amplification, without distortion, affect the intelligibility of the transmission. In other words, it shows the articulation of distortionless transmission for various degrees of loudness. The two scales along the abscissæ are, respectively, those of intensity and loudness in the units already described. Here, however, the zero level of each

has been shifted to that of the initial intensity or loudness of the sounds issuing from the lips of the speaker. The ordinates represent articulation in terms of the percentages of sounds correctly recorded.

Interpreting this curve we see that the initial intensity of undistorted speech emerging from the speaker's mouth may be increased a hundredfold or reduced to one-millionth of its value without seriously affecting intelligibility. The articulation remains above 85 throughout a range of which the maximum intensity is one hundred million times the minimum. It is interesting to note that as the intensity decreases from the initial value, the articulation actually becomes better, reaching a maximum at about one thousandth of that value. Even down to values of about one hundred-thousandth the articulation is as good or better than at the initial intensity.

Man's perception and understanding are, therefore, most efficient for the range of intensities to which he is most accustomed, for, of course, speech sounds usually reach the ear at intensities very much lower than when they leave the speaker's mouth. This is one of the many striking illustrations of how, by natural evolution throughout the ages, man's ear and understanding have developed to best meet the conditions under which he lives.

When the intensity is reduced to about one ten-billionth of its initial value, the articulation becomes zero, at which point also the sounds cease to be audible. Thus the threshold of hearing, coincides approximately with what we might call the "threshold of intelligibility." It is very important to observe, however, that below intensity values of about 10^{-6} or loudness value of -60 on the scales used in Fig. 76, the articulation falls off rapidly and becomes entirely unsatisfactory for very much smaller intensity or loudnes values. This points to one of the exacting requirements in the design of telephone systems. Even for undistorted transmission the intensity of the received sounds must be kept above a certain level in order that intelligibility may not suffer. Above this level a very wide range of intensities may be satisfactorily employed.

When the intensities of speech sounds are amplified very much above their initial values, the intelligibility is lessened, and there is a marked change in quality. Under these conditions the ear mechanism is overloaded. Subjective harmonics are

introduced so that even though the received sound waves are not distorted, there is a decided distortion of the sound as sensed. This interferes with both intelligibility and naturalness. If the sounds are still further increased in loudness, the threshold

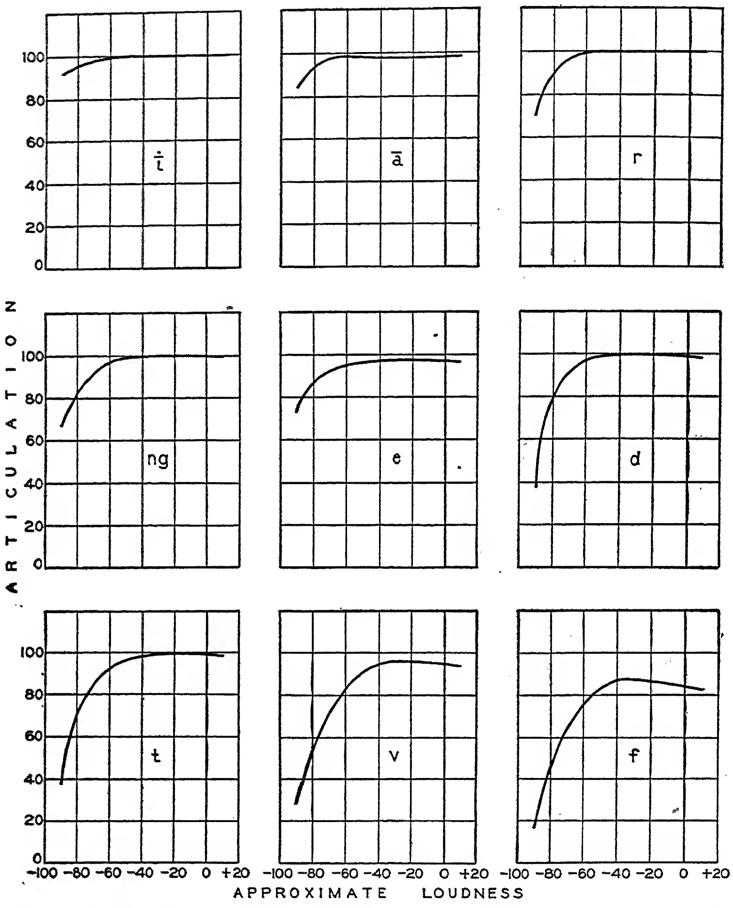


FIG. 77.—Articulation-loudness curves for some characteristic speech sounds. (*Courtesy of Bell Telephone Laboratories.*)

of feeling is crossed and a tickling sensation or actual pain is felt.

We may now inquire how changes in intensity or loudness affect the articulation of individual speech sounds. Figure 77 shows some of the results of investigations made in the Bell

Telephone Laboratories for a number of such sounds transmitted without distortion. As in Fig. 76, these curves show the percentage of times each of the sounds was correctly understood when received at various degrees of loudness. The general results of these tests are thus summarized in Mr. Jones' paper already referred to:

"It is observed that in general diphthongs and vowels are more easily heard than consonants, and that of the latter the stop consonants are heard with fewer mistakes than are the fricative ones. If all the sounds are listed in order of average articulation the top quarter will contain no consonants and the lower half no vowels. When speech becomes weak, the errors of the consonants increase greatly, their articulation values falling off at higher intensities than is the case with the vowels.

"There are some exceptions to these general statements. At moderate volume the short vowel *e* is near the bottom of the list, but at very weak volume 22 sounds are harder to perceive. *l*, *r*, and *ng* are all more readily heard than *e* at moderate volume, but when very weak they fall below it. *l*, which ranks with the diphthong *i*, as one of the easiest sounds at moderate volume, is mistaken about two times out of three when very weak.

"The diphthongs *i*, *ou*, and the long vowels, *o*, *ō*, *ā*, all have average articulations better than 95 and even when very weak have value of 84 or better. On the whole the sounds *th*, *f*, *s*, and *y* are hardest to hear correctly, and they account for more than half of all the errors of interpretation. In general, it is observed that the volume at which errors begin to be large is different for different sounds and is usually higher for the consonants than for the vowels. Within the precision of the data, the intersections on the axis of abscissas all correspond with the threshold of hearing."

The curves of Figs. 76 and 77 relate to practically distortionless transmission wherein all the frequencies are reproduced in their proper proportions, varying only the intensity or loudness of the whole complex sound. But, as has been shown, the maintaining of a proper loudness level of received speech is only a part of the problem of the sound-transmission engineer. In the interests of economy he must cut off some of the higher frequencies from his received speech and, therefore, another question to be answered is how far he may go in this direction without an undue sacrifice of intelligibility.

By introducing filters into the otherwise distortionless experimental system, certain bands of frequencies may be cut out of the received speech practically without interference with the others. Thus a filter may be introduced which will, in effect, cut off all frequencies above 1,000 cycles by reducing them to about one-millionth of their former value and yet scarcely affect the frequencies below 1,000. Such a filter which allows current waves below a certain frequency to pass while cutting off all frequencies above it is called a "low-pass" filter. Likewise, one which acts to pass high and to cut off low frequencies

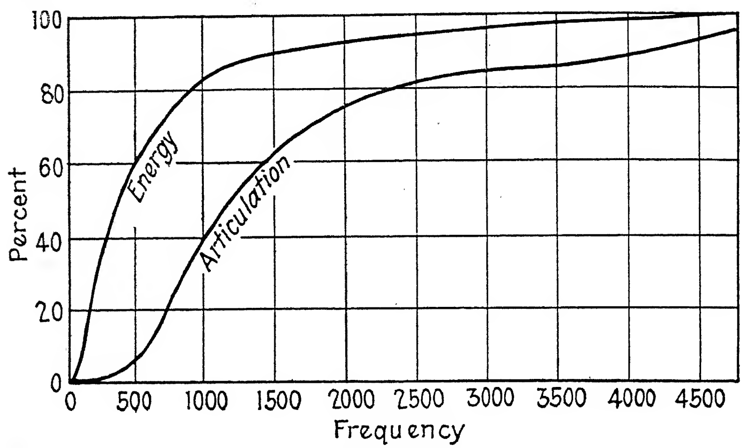


Fig. 78.—Effects on energy and intelligibility of cutting off frequencies above various points.

is called a high-pass filter. The frequency above or below which the filter discriminates is called the "cut-off frequency" or "cut-off point."

With this understanding the curves of Figs. 78 and 79 may be discussed. Figure 78 shows the effect on the articulation and on the energy of the received speech caused by cutting off all components *above* various frequencies. A low-pass filter is here used and its cut-off point adjusted to any frequency, as indicated along the abscissæ. Figure 79 shows similar effects on articulation and energy by the application of a high-pass filter which will cut off all frequencies *below* various values, passing only those above.

It is at once apparent from these curves that the frequencies below 1,000 cycles, while carrying most of the energy, are of

comparatively little value in contributing to good articulation. Thus Fig. 79 shows that if all frequencies below 1,000 are eliminated, the energy will be reduced to about 15 per cent of its former value, while the articulation will remain good at about 86 per cent. On the other hand, Fig. 78 shows that if we eliminate all frequencies above 1,000 cycles, we retain about 85 per cent of the energy and reduce articulation to 40 per cent. The results at the 500 cycle cut-off points are even more striking. Eliminating all frequencies above 500 ruins articulation but leaves 60 per cent of the energy, while cutting off the frequencies below this point scarcely affects the articulation but leaves only 40 per cent of the energy.

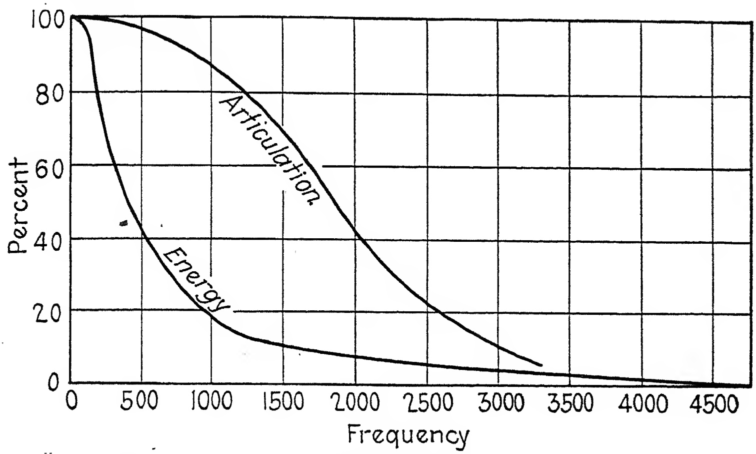


Fig. 79.—Effects on energy and intelligibility of cutting off frequencies below various points.

The two articulation curves of Figs. 78 and 79, if superimposed, would cross at about 1,550 cycles. This means that the articulation would be the same, about 65 per cent, if all frequencies either above or below 1,550 cycles were eliminated. The quality would be very different however, that obtained with the high-pass filter being described as "high and shrill," that with the low pass "low and dull."

It has been commonly thought that the range of speech frequencies from 60 to about 500 cycles, corresponding to the fundamentals and lower harmonics of the tones produced directly by the vocal cords, were the most important of all the speech frequencies. This is quite contrary to the facts. The funda-

mental vocal cord frequencies contribute very little to intelligibility. In fact, as shown by Fig. 79, the entire range of frequencies lying below 500 cycles may be entirely eliminated with scarcely any loss in clearness of articulation. There is, however, a decided change in the naturalness of the reproduced speech, much of its fullness or roundness of tone being lost.

Here we see the importance of the recently discovered facts, already alluded to, that the pitch of a tone is not always deter-

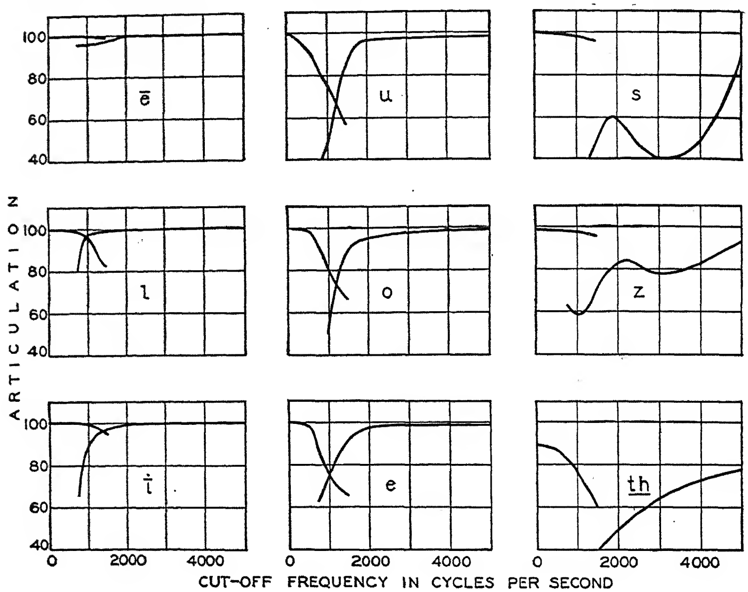


Fig. 80.—Articulation-distortion characteristics for some typical speech sounds. (*Courtesy of Bell Telephone Laboratories.*)

mined by the fundamental frequency of vibration, and that in the case of low tones it may be carried to the ear by the common difference between successive harmonics. Thus when the fundamentals and some of the lower harmonics are eliminated artificially from speech sounds, the ear, to a certain extent, supplies them subjectively.

Figure 80 shows the effects on the articulation of certain individual speech sounds due to suppressing components above or below various frequencies. Concerning these Mr. Jones says:

"The ordinate gives the number of times the sound was correctly observed per 100 times called; the abscissa, the fre-

quency of cut-off. In each figure the effect of suppressing the frequencies below the cut-off is shown by the curve at the left, the effect of suppressing those above it by the one at the right. The diphthong $\bar{\imath}$, the long vowel $\bar{e}$, and the semi-vowel l are each perceived with an error less than three per cent when either half of the frequency range is used. The intersections of the two curves, the cut-off frequency where the articulation is the same with either low-pass or high-pass filters, are at different points in each of the three cases, however.

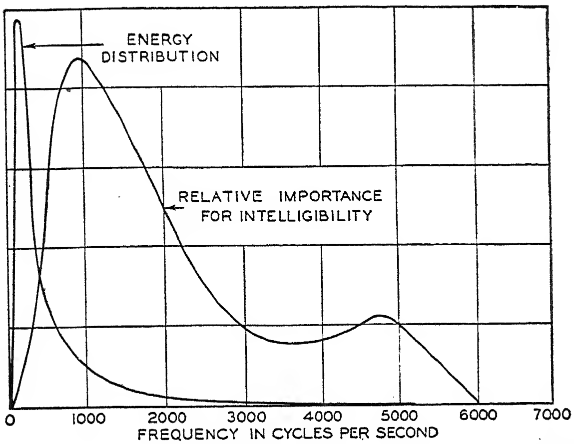


FIG. 81.—Frequency characteristics of speech. Curve *A*, energy distribution. Curve *B*, relative importance for intelligibility. (*Courtesy of Bell Telephone Laboratories.*)

"In the cases of the short vowels *u*, *o*, and *e*, the frequencies below 1,000 cycles are important to good articulation, but those above 2,000 may be suppressed with little effect.

"In the cases of the fricatives, *s*, *z*, and *th*, quite different effects are observed than with the former two classes. Some of the peculiar results shown have not yet been explained. Even if all frequencies up to 5,000 cycles are correctly transmitted, these sounds are noticeably impaired by the suppression of those above. The lower frequencies up to 1,500 cycles contribute practically nothing to the articulation of *s* and *z*. It has been observed, in the case of a system which suppresses all frequencies above 2,500 cycles, that about 82 per cent of the syllables were heard correctly in an articulation test, and that the errors were made up principally of failures in the three sounds *s*, *z*, and *th*."

The curves of Fig. 81[1] summarize in a general way some of the foregoing and other information that has been gathered concerning the frequency characteristics of speech sounds. By placing the two curves on the same diagram the dissimilarity between the energy distribution throughout the frequency range and the relative importance of the various frequencies as affecting intelligibility is emphasized.

The total speech frequency range extends from about 60 to 8,000 or perhaps even 10,000. Practically, however, the overall range does not extend above 6,000, and we may consider that a system which could transmit and reproduce all frequencies up to 6,000 without distortion would be capable of perfect transmission. As stated before, however, it is, at present, impracticable to design telephone systems to transmit so wide a band as this, nor is it necessary. It is found that a telephone transmission system, particularly one where the lines are loaded, acts, in effect, as a filter, transmitting most effectively a certain band of frequencies and suppressing others. Obviously, the transmission design should be such as to pass the most important band.

It will be seen from curve *B* of Fig. 81 that the most important range, so far as affecting intelligibility, lies quite distinctly below 3,000 cycles and this is now about the "cut-off frequency" that is aimed at in transmission design. Formerly, a system was considered satisfactory if it transmitted frequencies up to about 2,200 cycles, but recently the standard has been raised in order to obtain the betterment in articulation afforded by the added frequencies up to 3,000.

As to the lower end of the frequency spectrum, we have seen from the curves of Figs. 78 and 79 how little the frequencies below 500 cycles contribute to intelligibility. Their importance from the standpoint of naturalness of reproduction, however, is not to be ignored. Practically satisfactory speech transmission is secured if the lower end of the frequency spectrum is cut off at 300 cycles.

[1] MARTIN, W. H. and HARVEY FLETCHER, High Quality Transmission and Reproduction of Speech and Music, *Journal of the American Institute of Electrical Engineers*, March, 1924.

Curve of Energy Distribution given in Analysis of the Energy Distribution of Speech, by Crandall and MacKenzie, *Physical Review*, vol. 19, No. 3.

Curve for Intelligibility derived from data given in The Nature of Speech and its Interpretation, by Harvey Fletcher, *Journal of the Franklin Institute*, June, 1922.

For the practical purposes of commercial wire telephony, therefore, we may consider the range of transmitted frequencies to be from about 300 to 3,000 cycles. For very high quality speech transmission, however, where loud speakers are involved, this range must be considerably extended so as to employ currents having frequencies of from about 100 to about 6,000 cycles.

Music.—At first telephony had little to do with the transmission of music, but of late, with the coming of radio broadcasting, the addition of music to the telephone load has become of increasing importance. With it has come new problems.

Music is built up of tones of various qualities separated by definite intervals of pitch. The tones are employed in such succession and combination as to produce melody and harmony. As stated before, we cannot completely exclude noise from music. A single tap on a drum or a single clash of cymbals is hardly musical in itself, yet such sounds play an important part in music. The firing of cannon has even been employed to heighten orchestral effects.

The requirements for the transmission of music are more exacting than those for speech. The transmission of music must not only be of the kind that would give intelligibility in the case of spoken words or song but also it must be natural. It must be a faithful reproduction of all the sounds constituting the music because music transmission is for the purpose of entertainment and enjoyment, and, unless the natural quality of the original music is preserved, much of the enjoyment is lost and perhaps pain substituted.

The average radio loud speaker up to about 1926 illustrated this distinction between intelligibility and naturalness. Often they would reproduce speech with quite satisfactory intelligibility but hardly with naturalness. The words and sense were clearly understood but it was often impossible to recognize who was speaking. The same loud speaker, however, except in the opinion of its amateur owner, made a dismal failure of music.

The reason for this is instructive in its bearing on the general problem of sound transmission and reproduction. The improperly designed loud speaker horn is incapable of taking up and reproducing vibrations of frequencies lower than about three hundred. The elimination of these frequencies does not seriously impair the intelligibility of speech but it does take much of the

richness out of music. Again the radio or phonograph horn, faulty in acoustic design, is more responsive to some frequencies than to others, due to its own resonance. Thus its efficiency differs greatly for different frequencies, resulting in the accentuation of some tones at the expense of others. The resulting distortions may or may not seriously affect intelligibility of speech, but they always interfere with the quality of music.

A striking argument for the study of acoustic science and for the practical application of its principles is found in the very great improvements in the quality of musical reproduction by phonographs and by radio since about 1926. Prior to that time the development of even so simple a device as the phonograph horn had been largely by cut-and-try methods. The application of really correct acoustic principles has revolutionized its performance. Similar application of proper acoustic design has shown that it is possible to produce an electrical transmission system so perfect in all its stages from transmitter to final loud speaker that the received sounds, whether of music or speech, are indistinguishable from the original. If this is now true in the laboratory, it is not too much to expect its more general commercial realization, at least in the realm of music. It must be borne in mind, however, that this is a distinct field from ordinary commercial telephony. It is doubtful if it will ever pay to design the wire systems of communication for the more exacting requirements necessary for the most effective transmission of music. While music transmission over wire lines is sometimes desirable, it is of little importance in comparison with speech transmission.

Aside from the human voice the principal instrumentalities employed in producing musical tones are vibrating strings, air columns and membranes. Stringed instruments are exemplified in the violin family, the piano and the harp. Those employing air columns are the pipe organ, the various horns, clarinets and flutes. Those employing membranes or tympani are the drums and the cymbals. As has been shown, the tones of these instruments, even though of the same pitch, differ in quality or timbre according to the number and strength of their respective overtones.

Figures 82, 83, 84 and 85[1] each show the wave form and corresponding spectrum of component frequencies for the piano,

[1] The Physical Properties of Speech, Music and Noise, by Harvey Fletcher, presented before the New York Electrical Society, February, 1924.

violin, cello organ pipe and trombone organ pipe sounded at the respective pitches indicated. Similar diagrams for the clarinet were shown in Figs. 50 and 51 of Chap. V.

Figure 86[1] shows the continuous sound spectrum for organ music, the instrument in this case evidently having no pipe lower

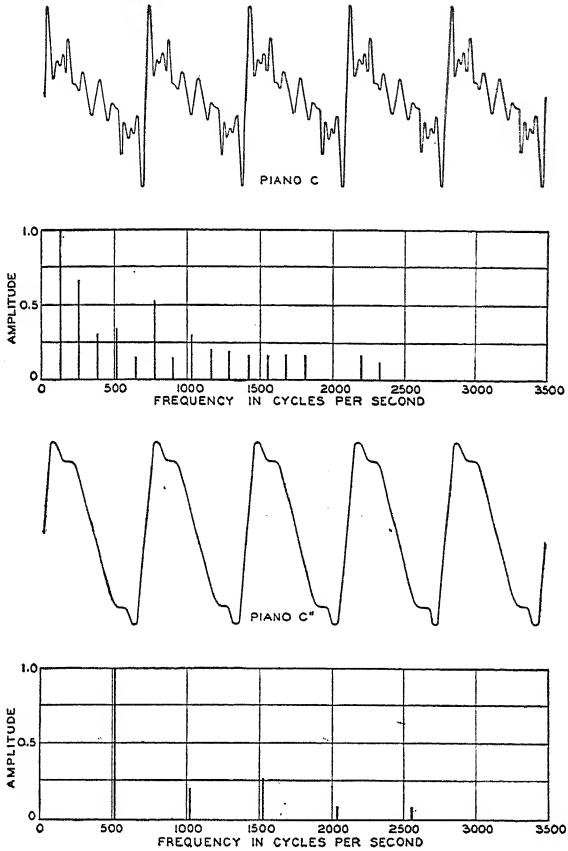


Fig. 82.—Wave forms and component frequencies for piano c and c''. (*Courtesy of Bell Telephone Laboratories.*)

than C_1. A consideration of this diagram will show why it is difficult properly to reproduce organ music telephonically. Neither the electrical system nor the receiver instruments are ordinarily designed for the very low frequencies, which, as the diagram shows, predominate.

The sounds of musical instruments also differ among themselves with respect to their duration. Certain instruments, like the

violin, horn and organ, are capable of long sustained tones while the percussive instruments, like the drums, produce only sudden sounds, sustained effects being produced only by rapid repetition

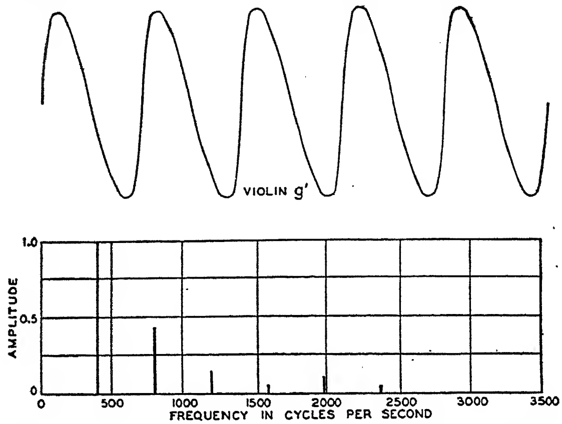


FIG. 83.—Wave form and component frequencies of violin g'. (*Courtesy of Bell Telephone Laboratories.*)

of the single percussions. The piano, while a percussion instrument, is capable, in its lower register, of producing sustained tones of moderate duration, due to the continued vibration of the string after the percussive blow.

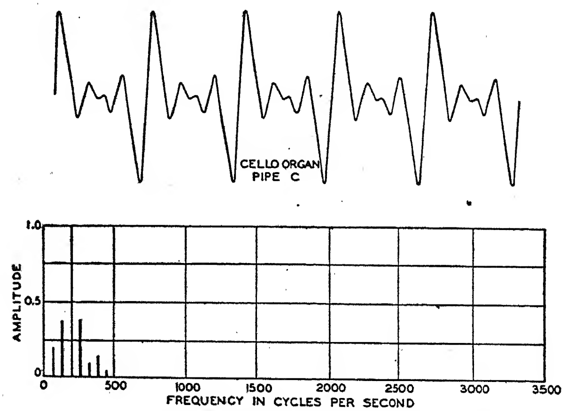


FIG. 84.—Wave form and component frequencies of organ-pipe C. (*Courtesy of Bell Telephone Laboratories.*)

As shown in Chap. V, the fundamental frequency range of the piano extends from about 27 to slightly over 4,000 cycles. Organ pipes go somewhat lower. This is not the entire range of useful musical frequencies because of the upper harmonics

associated with the fundamentals. The useful range of musical instruments, including harmonics, extends from about 16 to well above 10,000 cycles.

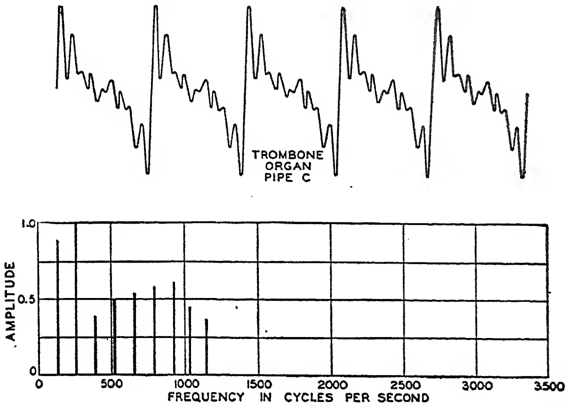


Fig. 85.—Wave form and component frequencies of trombone organ-pipe c. (*Courtesy of Bell Telephone Laboratories.*)

Here again, even in high quality transmission, it is not feasible to attempt the reproduction of the entire band of frequencies. Good reproduction of most kinds of music can be attained with a frequency range of from 50 to 5,000 cycles. Strange as it may seem, the low tones of the organ or piano can be heard in music

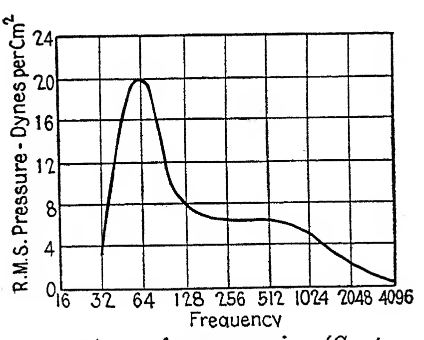


Fig. 86.—Continuous spectrum of organ music. (*Courtesy of Bell Telephone Laboratories.*)

reproduced by a system which electrically and acoustically is incapable of transmitting frequencies below 50 or 100 cycles. The ear senses tones lower than any actually present, supplying subjectively the suppressed fundamentals from the intervals between their successive harmonics above.

But it is only partially because of its wider frequency range that the transmission of music, with proper regard to quality, is much more difficult than that of speech. Another thing which adds to the difficulty is the enormous range of intensities encountered in ordinary orchestral effects. In these the ratio of the loudest to the softest tones otherwise alike may be as much as 100,000:1. For speech the corresponding intensity range is nearer 1,000:1.

Noise.—The sources of noise are almost infinite in variety and number. Of the three characteristics of sound—pitch, loudness and timbre—noise, according to the definition already given, has no recognizable pitch. Its loudness, however, is quite as readily recognizable as in other kinds of sound. As timbre or quality of a complex tone depends on the blending of its fundamental and various overtones, each of which has its own pitch, it follows that the ear, recognizing no pitch in noise, will have difficulty in recognizing quality.

Nevertheless, noises do have certain attributes other than loudness, which enable the ear to distinguish among them. The falling of a book and of a dinner plate on a stone floor would hardly sound alike, even though of the same loudness. The sound caused by the book would be in the nature of a thud, that of the plate a crash. Neither would have a recognizable pitch although one sound would be characterized as sharper than the other. The book would not break into small pieces, while the plate probably would, thus causing a large number of concomitant sounds due to the breaking and falling of the pieces. Moreover, it is likely that the plate would give forth a ringing sound, which could be properly regarded as tone and which would lend character to the crash of the plate not found in the thud of the book. Thus, noises may have certain attributes of timbre caused by the presence of recognizable tones added to the general confusion of vibration. They may also differ with respect to the concomitant sounds involved in their make-up and also in the degree of suddenness with which they occur. Only in a general way, therefore, may we assign quality or timbre to noise, and, in so far as we do assign it at all, it is probably due to the presence of recognizable tones which the ear can distinguish from the general confusion of vibrations.

The frequencies involved in noise probably cover the whole range of audition. Often noises are characterized by irreg-

ularity and suddenness. For these reasons, and in general because of their disordered make-up, noises are more difficult to transmit than the better ordered varieties of sound. Fortunately, however, it is seldom desired to transmit them. The chief consideration is to find means of suppressing them.

From the telephone standpoint, we may divide noises into two classes: room noises and line noises.

Room noises are the ordinary noises existing in the rooms at the transmitting and receiving stations. The noises of typewriters, of rattling paper, of footsteps, the babel of voices and the sounds coming into the room from street traffic are common examples.

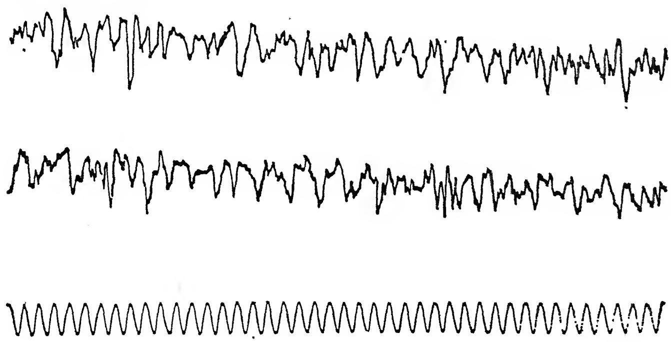


Fig. 87.—Typical wave form of a street noise and a pure tone of 500 cycles. (*Courtesy of Bell Telephone Laboratories.*)

In Fig. 87 is shown an oscillographic record of street noises. The absence of any regularly recurring wave form or of any fixed rate of vibration is noticeable. For purposes of comparison the record of a pure tone of 500 cycles per second made on the same scale is given at the bottom of this figure. Some idea of the complexity of street noises and sounds may be gained from the fact that the record here shown covered only about one-tenth of a second, and from the further fact that the oscillograph recorded only the major waves, undoubtedly omitting the smaller fluctuations due to the higher frequencies.

Line noises are those produced in telephone receivers by currents other than those due to the transmitter. As a matter of fact, these sounds are more likely to be tones than noises under the definitions just given. They are extraneous to the sounds it is desired to transmit, always objectionable, and, therefore, classed by the telephone man as "noises."

Line noises may be due to a number of causes. The principal ones are induction from neighboring power lines and irregularities in the direct current flowing from the generators which charge the battery in the telephone central office. Other line noises are caused by induction from neighboring telegraph or telephone lines. These latter causes are less troublesome because the inducing currents are relatively small. Where the line noise is due to induction from telephone currents in other telephone lines, it is called "cross-talk."

Line noises due to induction from power lines will have a characteristic tone corresponding to the frequency of the power line, usually 60 cycles per second.

In an artificially produced example of line noise, analyzed by Fletcher, all the odd harmonics of the 60-cycle fundamental were measurably present up to the fifty-fifth. This corresponded to a range of frequency from 60 up to 3,300 cycles, thus extending a little below and above the range of frequencies most used in speech transmission.

Room noises cover a much wider range and, under average conditions, their energy is found to be scattered with approximate uniformity throughout the entire audible range of frequencies.

The elimination of extraneous noises, whether line or room, is of more than passing interest. It may be characterized as one of telephony's outstanding problems. This may be more readily appreciated by reverting to the phenomenon of masking. In connection with Fig. 70 of Chap. VI it was shown how about half the length of the basilar membrane might be thrown into vibration by a single loud impressed tone. No additional sensation would be produced by other impressed sounds unless these were loud enough to cause greater amplitudes of vibration than those already existing.

With this in view we may consider the effect of line or room noise on the sensitivity of the ear in receiving transmitted sounds of speech. Figure 88 is given to illustrate the probable amplitude of response along the basilar membrane to a loud telephone line noise. Almost the entire length of membrane is thus "appropriated" by the useless line noise and the desired speech sounds cannot be heard until they are raised to a level loud enough to cause greater amplitudes of vibration than those shown in the figure. It is for the same reason that one has to shout in noisy places in order to be heard by a person of normal hearing.

All extraneous noises, whether line or room, tend to interfere with the reception and understanding of the sounds being transmitted. Only by making the received sounds loud enough to override the noise can they be heard. This means that a far higher standard of intensity of the received sounds must be

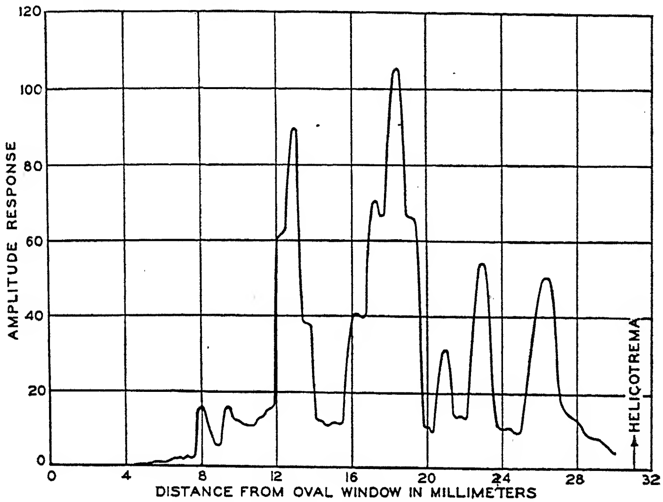


Fig. 88.—Probable amplitude of response of basilar membrane for a loud telephone "line noise." (*Courtesy of Bell Telephone Laboratories.*)

maintained than would be the case if complete quiet prevailed. To maintain this higher standard is costly, particularly with respect to the character of the plant required. Fletcher goes so far as to state that if in some way all the noises could be eliminated during the reception of speech, the present reproduced intensity of speech could be reduced to less than one hundredth of its present value.

CHAPTER VIII

VOICE CURRENTS

We have seen in earlier chapters that in electric telephony we do not actually transmit the sound waves, but rather, we convert them into corresponding electric current waves which are sent over the line to the desired point and there converted again into sound waves. These fluctuating electric currents which correspond to the sound waves we may call "sound currents," but since the sounds involved are usually those of the voice, we may adopt the somewhat narrower term, and refer to them as "voice currents." These are the current waves which Bell in the famous fifth claim of his patent characterized as "electrical undulations, similar in form to the vibrations of the air accompanying the said vocal or other sounds." It is with the characteristics of these electrical undulations or voice currents, and with their production and control, that this chapter will deal.

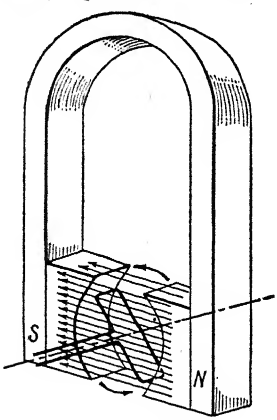


Fig. 89.—Generating sine wave of alternating electromotive force.

In Chap. V it was shown how sound waves in air, no matter how complex could be resolved into a number of simple harmonic wave motions of proper amplitudes, frequencies and phase relationships. Evidently then, since the voice current waves are to conform exactly to the sound waves, they may be subject to exactly the same sort of analysis. Fourier's theorem is just as true of current waves as of sound waves. We may say then, that voice current waves, no matter how complex in form, may be resolved into a number of simple harmonic or sinusoidal current waves of proper amplitudes, frequencies and phase relationships.

In connection with Fig. 29 (Chap. V) we considered the uniform motion of a point in a circular path in gaining fundamental

conceptions of simple harmonic motion and of a simple sound wave. In the same manner we may consider the uniform rotation of a conductor in a uniform magnetic field in arriving at fundamental ideas of a simple alternating current wave.

The loop of Fig. 89 may be considered as a part of the winding of an elemental dynamo. Its rotation about its axis is at a uniform velocity, in the direction indicated by the arrow, and the field in which it revolves is of uniform intensity throughout its entire cross-section. It is evident that as the loop revolves the portion of the conductor parallel to the axis will cut through the lines of force, first in one direction and then in the other, thus producing electromotive forces correspondingly alternating in direction. If the circuit of the loop is closed, corresponding waves of current will flow, successively alternating in direction.

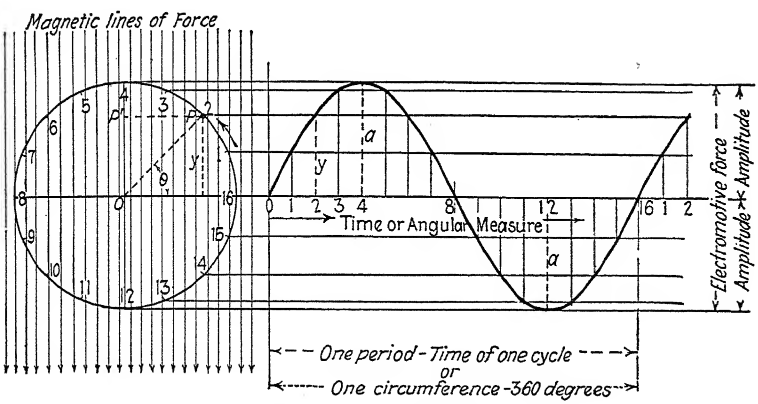


FIG. 90.—Sine wave of electromotive force.

We may best analyze the nature of these current waves by considering Fig. 90. Here the point P is an end view of one of the sides of the revolving loop, and the assumptions as to uniform circular rotation and density of field are the same as before.

The fundamental law to keep in mind in making this analysis is that when a conductor is so moved as to cut across the lines of force in a magnetic field an electromotive force will be set up in it, the magnitude of which will depend on the rate of cutting, and the direction of which will depend on the direction of cutting.

With this law in mind it is evident that as the conductor passes the point O it will be traveling upward, parallel with the lines of force and not cutting them at all. The electromotive

force in it will, therefore, be zero. As its movement progresses it will begin to cut through lines and, although its angular velocity is uniform, its rate of cutting lines will increase until at the point 4 it will be moving horizontally and cutting lines at a maximum rate. At this point the electromotive force will be at its maximum. Through the next 90 degrees of rotation, the rate of cutting and the generated electromotive force will decrease, until, at the point 8, 180 degrees or a half cycle from the start, the conductor will again be running parallel to the lines of force rather than cutting through them. The electromotive force will, therefore, have decreased to zero. For the next half cycle the rate of cutting and the electromotive force, in the same manner, will increase to a maximum at point 12 and decrease again to zero at point 16 where the cycle starts over again. On the second half cycle the electromotive force will be in the opposite direction from that in the first, because the conductor is cutting through the lines in the opposite direction.

We see, then, that in passing once around the circle (one cycle) the electromotive force passes continuously from zero to a positive maximum, then through zero and to a negative maximum and, finally, to zero to begin another cycle. The curve showing the moment-to-moment values of the electromotive force wave throughout this cycle may be plotted as at the right of Fig. 90. It can be shown that the *rate* of cutting and, therefore, the value of the electromotive force generated at any instant is proportional to the value of the ordinate y at that instant. It is only necessary, therefore, to coordinate the values of y with the corresponding values of the angle θ, measured along the horizontal axis, or with the corresponding moments of time measured along that axis. The curve drawn through these successive coordinate points will show the form of the electromotive force wave and indicate the relative momentary value of the electromotive force at any instant of time or at any phase of the angular progress.

The form of this curve is easily recognized as identical with those of the various curves in Chap. V, representing the form of simple sound waves in air, or, in a broader sense, the displacements, at any moment of a point vibrating with simple harmonic motion. That the electromotive force curve of Fig. 90 is truly sinusoidal is apparent when it is considered that any ordinate y is directly proportional to the sine of the angle θ through which

the point P has passed at the corresponding instant of time. In fact, if we consider the circle as of unit radius then

$$y = \sin \theta.$$

Of course there are many other ways of generating a sinusoidal wave of electromotive force or of current than by the uniform rotation of a conductor through a uniform field of force. This particular method has been chosen for discussion here in order to again bring out clearly the relationship between the sinusoidal wave and uniform circular motion. This relationship affords the most convenient means of reasoning about sinusoidal wave phenomena, whether the phenomena be related to sound, electricity or other branches of science, and whether the reasoning be done mathematically or otherwise.

We may inquire very briefly concerning a few of these relationships, and into the interpretations of such a diagram as that of Fig. 90. As this diagram was developed it had reference to the cyclic generation of electromotive force in the conductor P as it revolved in its circular path through a uniform field. The curve at the right considered in this light is, therefore, an electromotive force curve. Obviously, however, if the conductor P was included in a circuit containing only resistance and subject to no external disturbing forces, the current would vary in exact accordance with the momentary variation of the electromotive force, and the curve would be equally useful as indicating the moment-to-moment variations of the current. In like manner it may be conceived as indicating the momentary values of any variable which changes in a sinusoidal manner with the passage of time.

Coming back to the original conception of the curve of Fig. 90, it is easily shown mathematically that in such a uniform circular motion of a conductor across a uniform field the rate of cutting the lines of force and, therefore, the electromotive force generated is exactly proportional to the sine of the angle θ, through which the conductor has passed since the beginning of the cycle. This means that the electromotive force is a sine function of the angle θ. This may be expressed

$$e = a \sin \theta.$$

where e is the instantaneous value of the electromotive force and a its maximum value. Obviously, a is the amplitude

of the electromotive force wave and is represented by the radius of the circle of reference.

We might, of course, refer to the velocity of the rotating conductor in terms of linear measure, in feet per minute for instance, but since we are dealing with uniform circular motion it is more convenient to refer to it in terms of angular velocity, as so many radians per second. We have then, $\omega = \theta/t$ or $\theta = \omega t$, where ω is the constant angular velocity in radians per second, and θ the angle described by the point in the time t. Using this value of θ our electromotive force equation then becomes

$$e = a \sin \omega t,$$

which means that the electromotive force is also a function of time.

We may also express this equation in terms of the period T, which is the time required for one cycle; or of the frequency f, which is the number of cycles occurring in one second.

Since the rotating point traverses 2π radians in traveling one circumference, the angular velocity $\omega = 2\pi/T$ and the electromotive force equation becomes:

$$e = a \sin 2\pi t/T.$$

Since the frequency $f = 1/T$, it may also be expressed

$$e = a \sin 2\pi f t.$$

These are some of the fundamental relationships of harmonic variation. As illustrated here, the variable was electromotive force, but the relationships are equally applicable to any other harmonic variable, as, for instance, the current flowing in a circuit or the displacement of a particle agitated by a sound wave.

Let us now consider a few of the fundamental facts regarding the flow of currents in electrical circuits. The subject naturally divides itself into two classes of phenomena: Those concerned with continuous currents, on the one hand, and with variable currents, on the other. Under variable currents, of course, are included alternating currents which rapidly and periodically change both in intensity and in direction.

The fundamental laws governing continuous or direct-current phenomena are few and simple.

1. Ohm's Law.—The strength I of a continuous current is directly proportional to the electromotive force E in the circuit

and inversely proportional to the resistance R of the circuit. Algebraically, $I = E/R$ or $E = IR$ or $R = E/I$.

2. **Joule's Law.**—The heating power P of a current is equal to the product of the resistance R and the square of the current I. Algebraically, $P = I^2R$.

Since by Ohm's law, $E = IR$ Joule's law may be written $P = I^2R = I(IR) = IE$. This will serve to calculate the power expended in a circuit, whether it be by heating resistance or by some other means, such as driving a motor or the storage of chemical energy.

3. **Kirchhoff's Laws.**—*a*. In any branching network of conductors the algebraic sum of all currents flowing toward a distributing point is zero.

This is the same as saying that as much current must flow away from a point as toward it. This must obviously be true under the steady state of flow which characterizes continuous currents.

b. The algebraic sum of all electromotive forces in a closed circuit is zero. Here the electromotive force consumed by the resistance (IR) is, of course, considered as a counter or negative electromotive force.

For variable currents and, particularly, for alternating currents, these laws do not apply so simply. The fundamental reason for this is that with continuous currents the conditions within the circuit have been reduced to a steady state, while with alternating currents they are in a state of rapid change. With continuous currents all energy is being consumed within the circuit as it is generated and at a steady rate. Consequently, there are no changes taking place which will react on the circuit from without. With alternating currents, on the other hand, the changes going on within the circuit react on conditions outside the circuit, which, in turn, react on the circuit itself. Energy does not flow in a steady stream but is alternately stored up outside the circuit and then given back to it.

Some of the differences between alternating- and direct-current phenomena may be stated with reference to the three fundamental continuous current laws set forth above:

1. **Ohm's Law.**—For alternating currents, the current for a given electromotive force is no longer wholly dependent on the resistance of the circuit. Instead, the current is dependent on the impedance Z, and the law becomes $I = E/Z$, instead of

$I = E/R$. The impedance is a rather complex quantity and is not wholly an attribute of the circuit itself. It has two components, the *effective resistance,* made up of the true ohmic resistance and of other energy consuming elements of the circuit, and *reactance,* which does not consume energy but merely represents its surging from one part of the circuit to another. Calling the effective resistance R and the reactance X, the way in which the two enter as components of impedance is expressed by the equation $Z = \sqrt{R^2 + X^2}$.

2. Joule's Law.—The equation $P = I^2R$ for the heat generated in a conductor by the flow of current through it retains the same form for alternating currents as for continuous, but care must be taken to express the current strength I in proper terms. At first thought it might be supposed that since the current varies from moment to moment its average value could be used in determining its heating effect. The average current for a full cycle of alternating current, however, is zero, there being just as much current in one direction as in the other. An ordinary direct-current ammeter measures average values and would register zero on an alternating current. The assumption is, therefore, clearly untenable, for, under it, I^2R would be zero, and we know from common experience with electric heating devices and incandescent lamps that alternating current has decided heating effect.

What then is the measure of an alternating current which determines its heating power? We know that for any instant the heat generated depends on the square of the current flowing at that instant. Considering successive instants in an alternating current cycle, it is evident that the average heat generated will depend not on the average of the instantaneous values of the current but on the average of the squares of the instantaneous values. Since the square of a minus quantity is always positive, it is evident that both the negative and positive halves of the current cycle will have positive heating effects.

In Joule's equation, therefore, $P = I^2R$, I^2 must be the average of the squares of the successive instantaneous current values and, consequently, I must be the square root of the average squares of the instantaneous current values. This value of an alternating current is commonly referred to as the *square-root-of-the-mean-square* value or merely as the *root-mean-square* value.

Because an ampere of alternating current measured in this way has exactly the same heating effect as an ampere of continuous current, the root-mean-square value is commonly also called the *effective* value. In view of this it will be apparent why alternating-current ammeters are sometimes of the so-called "hot-wire" type. They depend for their action on the heating effects of the current being measured and will, therefore, measure either direct or alternating currents with equal correctness.

Mathematically it is easy to show that the root-mean-square value of any sinusoidal function is equal to $1/\sqrt{2} = 0.707$ times the maximum instantaneous value. Thus a sine wave of alternating current or voltage which has a maximum instantaneous value of 1 ampere or 1 volt will have an *effective* value of 0.707 ampere or 0.707 volt.

If we are concerned only with the heat generated by the current in traversing ordinary resistance, then R, in Joule's equation, is merely ohmic resistance. If, however, other heating effects are involved, such as those due to magnetic hysteresis, dielectric absorption, eddy currents or various mechanical losses, then R must be taken as the total energy component of impedance, that is, the *effective* resistance to the current.

The power equation $P = EI$ applicable to continuous currents does not hold true for alternating currents, except where all the power is being consumed in heating the conductor. In this case the circuit contains only resistance and no reactance, and the equation holds true if the effective values of I and E are taken.

For circuits containing reactance, as will be shown, the current wave is not in phase with the wave of electromotive force that is impressed on the circuit. It may lag behind it or be in advance of it up to a maximum of a quarter of a cycle, or 90 degrees in either direction. As a result there may be large currents when the electromotive force is zero and *vice versa.* Again, the current may be negative for a positive electromotive force and *vice versa.* The product of E and I, therefore, means little, except for instantaneous values when it indicates only the power at the corresponding instant, sometimes negative and sometimes positive. For integrated values of power it can be shown mathematically that a third factor, the cosine of the angle of phase difference, must be applied. The power equation then becomes

$$P = EI \cos \theta.$$

The factor by which the direct-current value EI of the power must be multiplied to arrive at the alternating-current value is called the *power factor*. It is evident from the last equation that the power factor is equal to the cosine of the angle by which the current lags or leads the electromotive force.

As the cosine of an angle varies from 1 at 0 degrees to 0 at 90 degrees, it is apparent that for given current and electromotive force the power becomes less and less as the angle of lag or lead nears 90 degrees. Here the significance of the so-called "watt-less current" is seen, for, at a phase difference of 90 degrees, the power would be zero, even though very large current and electromotive force were present.

3. Kirchhoff's Laws.—As these apply to conditions of continuous-current flow they are not applicable in their ordinary sense to alternating currents. Continuous-current values may be directly added to and subtracted from each other. When a continuous current branches into several paths, the sum of the divided currents will always equal the undivided. The combination of two or more direct currents will always be their algebraic sum. This is far from true with effective values of alternating currents, for, with these, currents may divide into a number of components, each of which is larger than the undivided current, and, conversely, a number of large currents may combine into a smaller composite. The distinction is one of direction. Continuous currents have definite directions, either positive or negative. They may be treated as *scalar* quantities. Alternating currents, considered integrally, have no definite direction. In combining them their phase relationships must be taken into account. They must be treated as *vector* quantities.

It was stated above that these three fundamental laws did not apply *in such simple manner* to alternating- as to continuous-current phenomena. It would have been a mistake to say, as is so often done, that they do not apply to alternating currents, for, in the last analysis, they are equally true for both classes of currents. The difference is that in alternating currents they are more difficult to apply, because many additional considerations are introduced due to the alternate storing up and releasing of energy. The trouble in this respect with alternating currents is that, considered integrally, they have no direction and are changing from moment to moment. If, however, we consider the conditions *at any instant of time* we have both direction and

definite values. If, therefore, for any such instant, we take into account *all* the factors tending to cause or to oppose current flow, then these laws hold true for that instant in exactly the same manner as they do for direct currents.

Coming now to a somewhat more detailed discussion of alternating-current flow, we have seen that, for alternating currents, impedance takes the place of ohmic resistance, which, for a given electromotive force, limits the flow of continuous current. Also, we have seen that, broadly, impedance has two components: These are the apparent resistance R, which includes the ordinary ohmic resistance and all other energy consuming elements in or associated with the circuit, and the reactance X, which includes those elements which are wattless, *i.e.*, those which do not *consume* energy but merely alternately borrow and give it back as the current surges to and fro. Calling Z the impedance, these two factors enter into its composition thus:

$$Z = \sqrt{R^2 + X^2}.$$

The equation for alternating currents which takes the place of the simple expression of Ohm's law for direct currents then becomes:

$$I = \frac{E}{Z} = \frac{E}{\sqrt{R^2 + X^2}},$$

where I and E are, respectively, the effective current and the effective electromotive force.

Consider now the two components of impedance, R and X:

The effective resistance R is equal to the ohmic or continuous current resistance of the circuit only where all the energy is expended in heating the conductors. As a rule, however, in alternating-current phenomena energy is also expended outside the conductor which is not given back, and which acts, therefore, like ordinary resistance in the consumption of energy. These expenditures outside the circuit may be wasted or they may result in useful work.

The wasted energy or *losses* may be due to the following phenomena: Not quite all the energy required to magnetize an iron core is given back when the core is demagnetized (magnetic hysteresis); not quite all the energy required to charge a condenser is given back upon discharge (dielectric absorption);[1] eddy

[1] This has been commonly called "dielectric hysteresis" since it was thought to be somewhat analogous to magnetic hysteresis in iron. Recent

currents are caused to flow in any conductors which may be near the circuit (eddy current losses). At very high frequencies an additional loss, due to radiation of energy into space, becomes important. The useful expenditures outside the conductor are those which result in useful work, such as the moving of a telephone diaphragm or the development of mechanical power by an electric motor. Whether wastefully or usefully expended, none of the energy consumed outside the conductor in any of these ways is given back to the circuit, and their combined effect is to make the effective resistance R in the impedance equation larger than the ohmic resistance.

The other component of impedance, the reactance X, is the one which involves the surging back and forth of the energy in a circuit without its actual consumption or dissipation. This must be considered in two phases: First, the reactance due to electromagnetism, and, second, that due to condenser action or electrostatic capacity.

1. Reactance Due to Electromagnetism.—To facilitate an understanding of this we may briefly review a few fundamental principles of electromagnetic induction.

An electric current flowing through a conductor sets up a field of force about the conductor throughout its entire length. This field of force consists of magnetic lines extending in closed curves about the conductor and is sometimes termed a "magnetic whirl." A freely suspended magnetic needle placed within this field of force will tend to assume a direction corresponding to the direction of the lines of force and, therefore, at right angles to the conductor.

If the current flowing in the conductor is maintained at a constant value and in the same direction, the field of force about the conductor will not change. On the other hand, if the current strength fluctuates, the field of force will fluctuate, becoming more intense while the current strength is increasing and less intense while the current strength is decreasing. If the current changes

investigators have come to the conclusion, however, that *dielectric hysteresis* does not exist—that there is no exact parallel to magnetic hysteresis in a dielectric. The present tendency, therefore, is toward calling it *absorption.* Briefly the distinction is this: In magnetic hysteresis the loss per cycle is constant for a given maximum density. In "dielectric hysteresis" the loss per cycle is not constant but depends on the time required to complete the cycle.

its direction, the existing field of force is entirely destroyed and built up in the opposite direction. The building up of the field of force may be thought of as the originating of more magnetic lines of force about the conductor and the expansion of these outward, and the decrease of intensity as the collapse of lines of force into the conductor.

Whenever there is a relative movement between a conductor and the lines of force of a magnetic field, such as to cause the conductor to cut the lines or the lines to cut the conductor, an electromotive force is set up in the conductor which tends to cause a current to flow. The direction of this electromotive force will depend on the direction of the lines and on the direction of cutting, and its value will depend on the rate of cutting. The field of force may be set up either by a magnet or by a conductor carrying a current, and in either case the phenomenon just described is called *electromagnetic induction.*

If two wires, one of them carrying a varying current, are parallel and close together, then the lines of force set up by the current-carrying wire will, when the field of force contracts or expands, cut the second conductor inducing an electromotive force in it. If the two wires are formed into adjacent parallel coils, each having a number of turns, then some of the lines of the field of force set up by the coil carrying the current will cut some or all of the turns of the other coil. An electromotive force will thus be induced in each turn of the latter coil, the result being that the sum of all the electromotive forces induced in the separate turns will be added, thus producing a much greater effect than if the latter coil had but a single turn. The contracting or expanding of the field due to the coil carrying the current will take place only when the current in this coil is decreasing or increasing in value, and it therefore follows that electromotive forces will be induced in the second coil only when the current in the first coil is changing. This form of induction between one wire and another is *mutual induction.*

But consider only a single conductor: It is evident that if a single wire, coiled into many adjacent turns, is traversed by a current, each turn may be considered as setting up its own magnetic field, and that the field for any turn will to a greater or lesser extent link all the other turns. Any one turn will, therefore, have an inductive effect on all the other turns of the same coil when the current is varying. When the current decreases

the decreasing number of lines of force set up by this turn will act on all the other turns (as well as on itself) to induce in each an electromotive force tending to cause a current *in the same direction* as that already flowing. Similarly, the diminution of the original current in each of the turns will act on all the turns in the same way, and, as all of the electromotive forces so induced in all of the turns will be in the same direction as the current which is already flowing in the coil, their effects will be added and will tend to *prolong* the flow of current. On the other hand, an increase in the original current will cause an increasing number of lines of force to cut through all the turns, and this will induce electromotive forces in them tending to cause current to flow in the opposite direction to that already flowing, and, therefore, in a direction tending to *prevent* the increase in the original current. This phenomenon of induction between the turns of the same coil or between various parts of the same circuit is called *self-induction*.

In either of the phenomena just referred to, mutual induction or self-induction, the presence of iron in the magnetic circuit of the coiled conductors greatly increases the inductive effect. For this reason, where inductive effects are desired, the coils are usually wrapped about cores of iron and sometimes also entirely surrounded with iron. The reason for this is that a given magnetizing force, or force which tends to produce magnetic lines of force, will produce a far greater number of lines in iron than in air. The induced electromotive force is, therefore, greatly increased owing to the greater rate of cutting caused by changes in current.

In view of the fact that a decreasing current induces an electromotive force tending to produce a current in the same direction as that already flowing, while an increasing current induces an electromotive force tending to produce current in the opposite direction, it follows that the general effect of self-induction in a circuit is to tend to prevent any changes in current from taking place. The effect is clearly like that of inertia with respect to the movements of physical bodies. Self-induction accounts for the fact that coils of wire, such as those forming electromagnets, tend so greatly to retard the flow of rapidly varying currents, such as voice currents, through them.

Faraday's law of electromagnetic induction postulates that the electromotive force induced in a conductor by a varying magnetic

field is proportional to the rate at which the conductor cuts through the lines of force of the field. If the conductor is interlinked once with the field, it cuts through all the lines once when the field is created or destroyed, but if it is interlinked with the lines of the field n times, due to its being in the form of a coil with n turns, then the rate of cutting and, consequently, the induced electromotive force will be n times as great as for a circuit of one turn. Hence it is the total number of interlinkages between the lines of force and the turns of the coil that is to be considered in determining the rate of cutting.

The total number of interlinkages between the lines of force and the turns of the conductor is evidently the product of the number of lines threading the coil and the number of turns in the coil (assuming that all the lines thread all the turns). The number of turns is, of course, a characteristic of the coil, but other characteristics of the coil will also influence the total interlinkages that will be brought about by a given current. Principal among these are the character of the material in the magnetic circuit and its shape, cross-section and length.

If the coil is wound on an iron core, or if the magnetic circuit through which the lines of force pass consists in whole or in part of iron, a given magnetizing current will set up a much greater number of lines of force than if the magnetic circuit were entirely through air or other non-magnetic substances. This is because iron is much more *permeable* to magnetic lines than air or any of the non-magnetic substances. Again, a given magnetizing current will produce more lines in a short thick magnetic circuit than in a long thin one of the same material. In general the *reluctance* of a magnetic circuit, which is its resistance to the flow of magnetic lines through it, is proportional to the length of the circuit and, inversely, proportional to its cross-section and to the permeability of the material composing it.

It is evident, therefore, that the rate at which lines of force are cut by the conductor, and, therefore, the electromotive force of self-induction, will depend not only on the rate of change in the current flow but also on certain *characteristics of the circuit itself*. The term used to express the combined effect of these circuit characteristics with respect to their influence on the induced electromotive force which a given current will set up is the *coefficient of self-induction*. It is commonly referred to by the symbol L.

More explicitly, the coefficient of self-induction is the total number of interlinkages between lines of force and turns of the conductor (number of turns times total flux) that will be set up by a current of unit strength flowing in the coil. In other words, it is the total amount of cutting of lines by the circuit when a unit current is turned off or on, or when there is a change of one unit in the strength of the current. If ϕ is the total number of lines (flux) passing through the core of the coil and N the number of turns, then the total interlinking when the field exists or the total cutting when the field is created or destroyed will be ϕN, since each line will interlink with or cut through each turn. Then, if the flux ϕ is proportional to the current I producing it,

$$\phi N = LI \text{ or } L = \frac{\phi N}{I}.$$

The coefficient of self-induction is, therefore, the ratio of the total interlinking of the lines by the turns to the current producing them. This coefficient is a characteristic of the circuit[1] and its unit in the practical system of units corresponding to the volt, ampere and ohm is the *henry*.

The henry represents such an inductance as will result in the cutting of 10^8 lines of force in response to a change in current strength of one ampere. Since a conductor cutting through lines of force at the rate of 10^8 lines per second will have induced in it an electromotive force of one volt, it follows that a coil having a coefficient of self-induction of one henry is one in which a current changing at the rate of one ampere a second will produce in it an electromotive force of one volt.

Let us now inquire what actually happens, first, when a source of constant electromotive force and, second, when a source of

[1] It must not be thought that L is always a constant for a given coil. It is constant only for an "air-core" coil, or for any coil having only non-magnetic substances in its magnetic circuit. Here the magnetic field set up is strictly proportional to the current strength. For a coil having an iron core, however, or for one having iron in its magnetic circuit, the magnetic flux is not always exactly proportional to the strength of the magnetizing current. In other words, the *permeability* of the iron is not constant throughout the entire range of magnetization. For instance, as very high magnetic densities are reached the iron approaches "saturation," a condition where a further increase in current strength will have comparatively little effect on the number of lines through it. For very high magnetization, therefore, approaching saturation, the coefficient of self-induction of a coil is lower.

alternating electromotive force is applied to a circuit containing self-induction or inductance.

Considering the constant electromotive force first, we have, in Fig. 91, a diagram of a circuit containing a battery which we assume to have a constant electromotive force, a resistance R and an inductance L. For the purposes of discussion here R is supposed to contain *all* the resistance and L *all* the inductance of the curcuit. If there were no inductance the current would rise to its full value, $I = E/R$ the instant Key No. 1 was closed. With the inductance present, however, this cannot happen. A magnetic field must first be created by the coil, this requiring energy and time. In doing so, an electromotive force of selfinduction will be developed opposing that of the battery. This will be great at first but will rapidly disappear as the field approaches the full strength that the current will develop. Only when the field reaches its full strength, and, therefore, ceases to change, does the induced electromotive force disappear and allow the current to reach its full value, $I = E/R$, as determined by the resistance. We see then that the inductance in the curcuit, considered apart from the resistance, does not affect the final value of the current, but it does delay the time when that value is reached.

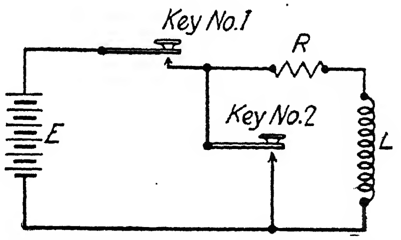


FIG. 91.—Circuit containing constant electromotive force, resistance and inductance.

Similarly, if the electromotive force is suddenly removed, as by short circuiting the battery with Key No. 2, the current will not at once drop to zero. The decrease in field due to the decreasing current will induce an electromotive force, this time in the same direction as that of the battery, thus for a time continuing the flow of current. Only when the decreasing field has ceased to change will the electromotive force of self-induction disappear and the current reach its zero value.

This rise and fall of current following the application and removal of a continuous electromotive force in an inductive circuit is shown graphically in Fig. 92. This brings out strikingly the fact, already mentioned, that in a circuit containing inductance the current does not rise or fall to its final values until a

time after the application or removal of the electromotive force which causes it to flow. Hence, we say that in inductive circuits the current values always lag somewhat behind corresponding changes in the impressed voltage.

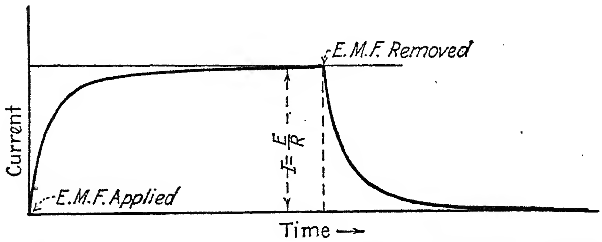


FIG. 92.—Rise and fall of current in a circuit containing resistance and inductance.

Considering alternating instead of continuous electromotive forces, we will take first a circuit assumed to have inductance only and find the relationship between an impressed sinusoidal electromotive force, the electromotive force of self-induction and the current. Such a circuit, indicated in the upper part of Fig. 93, could not actually exist, because the avoidance of all

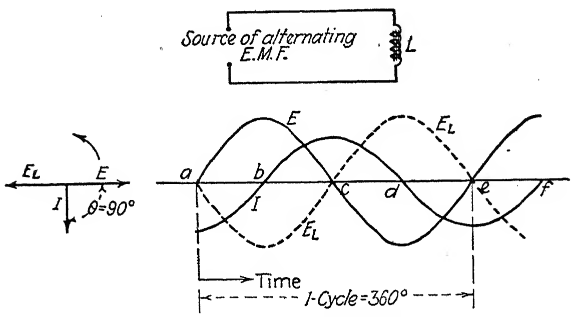


FIG. 93.—Current and electromotive force waves in circuit having inductance only.

conductor resistance is impossible. It is profitable, however, to analyze the action of such a circuit considered as a limiting condition. Here, there being no resistance in the circuit, the impressed electromotive force may be considered as applied directly to the terminals of the inductance.

Referring to the curves of Fig. 93, curve I represents the current. This current builds up a varying magnetic field whose changes

generate an electromotive force of self-induction, represented by curves E_L. At a the current is at its maximum negative value and, for the instant, is not changing. Accordingly, the value of E_L at that instant is zero. As the current I increases from its maximum negative value at a to its maximum positive value at c, the electromotive force of self-induction E_L will be negative in direction because opposing the increase of current. The current and, consequently, the field will be changing most rapidly at b and, therefore, E_L will have its maximum negative value at that time. During the next half cycle, from c to e, the current will be decreasing, and, consequently, E_L, tending always to oppose the change, will be positive and will reach its maximum value at d when the current is changing most rapidly.

An inspection of curves I and E_L of Fig. 93 shows that the electromotive force of self-induction E_L reaches its successive values always just a quarter of a period later than the current reaches its corresponding values. In other words the electromotive force of self-induction lags a quarter period or 90 degrees behind the current flowing. This must always be true of any inductance, whether there is resistance in the circuit or not, since it is the changing current which produces the changing field, and it is the changing field that produces the induced electromotive force.

Under the conditions of no resistance assumed, this electromotive force of self-induction E_L would be the only opposition offered to the impressed electromotive force. The resistance drop IR, always present under practical conditions, would be absent because, under the assumed conditions, $R = O$. Consequently, the impressed electromotive force must at each instant be equal and opposite to the induced electromotive force as indicated by curve E of Fig. 93. We have then, under the assumed resistanceless condition, a current I surging back and forth with no resistance to overcome and therefore wattless. Inspection of curves E and I shows that the current lags just 90 degrees behind the impressed electromotive force.

These three curves E, I and E_L, following each other in the order mentioned and separated by 90-degree intervals of time, may be considered to have been generated by the revolution of the three vectors bearing corresponding designations at the left of Fig. 93. As just stated, the angle between E_L and I must always be 90 degrees, regardless of whether there is resistance

in the circuit or not, but, as will presently be shown, the resistance in the circuit always reduces the angle θ, by which the current lags behind the impressed electromotive force, so that in practice it always is less than a right angle.

We may now take up a circuit representing more nearly practical conditions, such as is shown in Fig. 94. This circuit contains both resistance and inductance and we will assume a sinusoidal source of electromotive force to be acting. Here, as in Fig. 91, if the inductance were not present, the current would be opposed only by the resistance. In this case, however, the electromotive force is a varying one, and the current wave would rise and fall in exact accordance with the electromotive force wave. The current would be in phase with the electromotive force and the current equation would be, as in continuous currents, $I = E/R$.

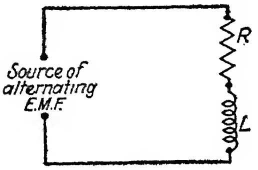


FIG. 94.—Circuit containing resistance and inductance.

With the inductance present, however, this simple state of affairs no longer exists. Following the variations of the current, a magnetic field is built up, first in one direction, then destroyed, built up in the opposite direction and again destroyed during each cycle. Energy is thus taken from the supply and restored to it twice during each cycle. The building up and destruction of this field, as already explained, develops an electromotive force of self-induction, and this together with the impressed electromotive force acts to drive the current through the circuit. It is important to inquire into the phase relationships of the two electromotive forces and the current that actually flows as a result of their combined driving force.

The current which actually flows in such a circuit as that shown in Fig. 94 will, as just stated, be due to the resultant of the impressed and the self-induced electromotive forces, both of which are sinusoidal. This current will also be sinusoidal, because the combination of two sine waves of the same frequency always results in a sine wave. The instantaneous values of this actual current may be represented by curve 1 of Fig. 95. This current curve may be considered as being generated by the vector OA revolving counterclockwise about the point O. The length of this vector will represent the maximum instantaneous current value I and will be the amplitude of the current curve.

Similarly, the electromotive force which drives this actual current will be the resultant of the impressed and the self-induced electromotive forces. It will be in phase with the current, and its instantaneous values are represented by curve 2 of Fig. 95. This curve of resultant electromotive force may be considered as being generated by the vector OB, coincident in direction with OA (since the two are in phase) and equal in length to the product IR of the current and the resistance.

It is the current actually flowing, curve 1, which causes the changes in flux which generate electromotive forces of self-induction. This electromotive force of self-induction will, as just shown, lag 90 degrees behind the current, and, accordingly,

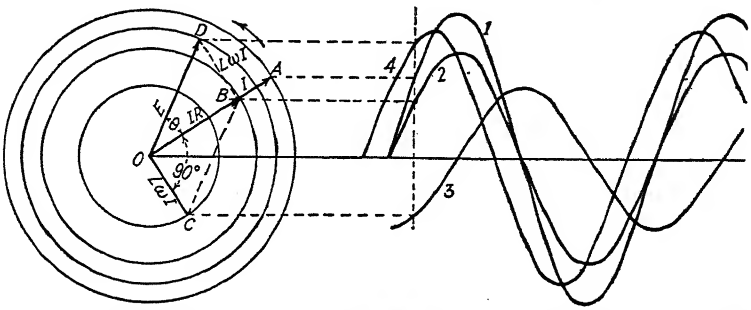


FIG. 95.—Current and electromotive force waves in circuit having resistance and inductance.

curve 3 of Fig. 95 may be taken to represent the instantaneous values of the electromotive force of self-induction. This passes through zero at the maximum and mimimum current values and has its maxima at times when the current curve is at zero. This curve may be considered as being produced by the vector OC, drawn 90 degrees behind OA and having a length corresponding to the maximum value of the electromotive force of self-induction.

Having now the resultant electromotive force represented by vector OB, and one of its components, the self-induced electromotive force represented by vector OC, we have only to complete the parallelogram of vectors, by drawing OD parallel to BC and DB parallel to OC, in order to determine the other component OD representing the impressed electromotive force. Evidently, curve 4 of Fig. 95, generated by revolving the vector OD about the point O, will represent the instantaneous values of the impressed electromotive force E.

It is now apparent that the actual current flowing in a circuit containing only resistance and self-inductance will lag behind the impressed electromotive force by an angular amount which may be measured along the horizontal axis of the curves of instantaneous value or, more directly, by the angle θ between the vectors. Obviously, this must always be less than 90 degrees since resistance must always be present in some degree.

The maximum value of the electromotive force of self-induction, needed to determine the length of the vector OC in Fig. 95, may be determined from the following considerations: Since the maximum value of the current actually flowing is I and the inductance is L, then at the moments of maximum current the total interlinkage of flux with the circuit will be LI. This flux will be created and destroyed twice during each cycle, so that the total cutting of lines in one cycle will be $4\,LI$. Since, for frequency f, there are f cycles in 1 second, the total cutting in 1 second will be $4fLI$. This is the average rate of cutting, and it can be shown that for sinusoidal variation the maximum rate is to the average rate as π is to 2. Hence, the maximum rate of cutting is

$$\frac{4\pi fLI}{2} = 2\pi fLI = L\omega I,$$

where

$$\omega = 2\pi f$$

is the angular velocity.

As it is the rate of cutting that determines the electromotive force induced, we may write

$$E_L = 2\pi fLI = L\omega I,$$

where E_L is the maximum electromotive force of self-induction and I the maximum current. Obviously, this is the required length of the vector OC of Fig 95.

The last equation may also be written:

$$I = \frac{E_L}{2\pi fL} = \frac{E_L}{L\omega}.$$

Here, the term $2\pi fL$ or $L\omega$ occupies the same position in expressing the relationship between the current and the electromotive force of self-induction as does the simple resistance R in Ohm's equation $I = E/R$ for continuous currents.

This quantity $2\pi fL$ or $L\omega$ is called the *inductive reactance.* It may be indicated by the symbol X_L, the subscript L being used

to show that this particular reactance X_L is due to self-induction, as distinguished from another kind of reactance X_c which is due to capacitance and which will be considered later.

Inductive reactance represents what may be called the "choking effect" of self-induction. It is to be noted that it is proportional to the frequency, which is a characteristic of the current, and to the coefficient of self-induction, which is a characteristic of the circuit. It is obviously independent of the resistance of the circuit.

The relationship of electromotive forces in an alternating-current circuit containing resistance and self-induction (Fig. 94) may always be shown by a right triangle such as OBD of the vector diagram of Fig. 95. Such a triangle of electromotive forces, stripped of other details, is shown in Fig. 96, this being taken directly from Fig. 95. The three sides of the triangle represent by their lengths the values of the respective electromotive forces and by their directions (vectorially speaking) the directions in which these electromotive forces are acting. Obviously, it will always be a right triangle, because the electromotive force of self-induction is always 90 degrees behind the actual current flowing, which, in turn, is in phase with the resultant electromotive force. Obviously also, the angle θ between the impressed and the resultant or active electromotive force will always represent the angular distance that the current lags behind the impressed electromotive force.

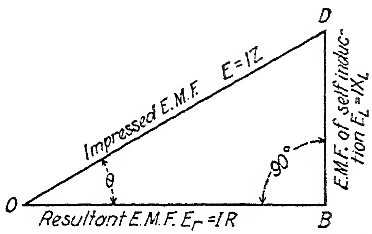


FIG. 96.—Electromotive force triangle for circuit having resistance and inductance.

This angle θ is called the "angle of lag." From the right triangle DOB of Fig. 95 or Fig. 96 it is seen that

$$\tan \theta = \frac{DB}{OB} = \frac{L\omega I}{IR} = \frac{L\omega}{R} = \frac{2\pi f L}{R}.$$

That is, in a circuit containing only resistance and inductance, the current will lag behind the impressed electromotive force by an angle whose tangent is

$$\frac{L\omega}{R} \text{ or } \frac{2\pi f L}{R} \text{ or } \frac{X_L}{R}.$$

It is thus seen, both from this formula and from an inspection of the diagram, that if the self-induction becomes zero the angle θ will be zero. In this case there will be no lag and the current will be in phase with the impressed electromotive force. On the other hand, if the resistance becomes zero, which was the condition assumed in Fig. 93, the tangent of θ will be infinite, which means that θ will be a right angle. This means that in a circuit containing inductance but no resistance the current would lag exactly 90 degrees behind the impressed electromotive force. Such a circuit is, of course, impossible of attainment and is to be considered as a limiting condition, as the ratio of the inductance to the resistance becomes very large.

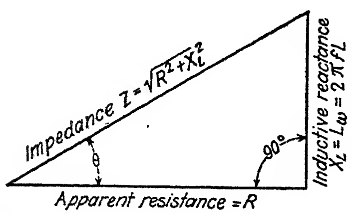


FIG. 97.—Impedance triangle for circuit having resistance and inductance.

If we divide the numerical value of each side of the electromotive force triangle of Fig. 96 by I, we obtain the exactly similar triangle of Fig. 97, which may be called the "impedance triangle." Each of its three sides shows the opposition offered, respectively, by the apparent resistance R, the inductive reactance X_L and the impedance Z; and also the direction in which each acts in exerting its opposition. It was stated earlier in this chapter that the impedance Z, which for alternating currents took the place of the resistance R as the limiting factor of the flow of current, was made up of two components, the apparent resistance R and the reactance X, and that these components were combined as indicated in the equation

$$Z = \sqrt{R^2 + X^2}.$$

The inductive reactance just discussed is the second component X of impedance in all cases where capacity effects do not enter into the circuit. It is apparent from this triangle why the inductive reactance X is not added directly to the resistance R to make up the total impedance offered by such a circuit. Since the inductive reactance acts at right angles to the resistance in opposing current flow, the combined effect of the two is represented by the hypotenuse of the triangle of which they are the two sides. As the hypotenuse of a right triangle is equal to the

square root of the sum of the squares of its two sides, it is evident that

$$Z = \sqrt{R^2 + X_L{}^2} = \sqrt{R^2 + L^2\omega^2} = \sqrt{R^2 + 4\pi^2 f^2 L^2}.$$

The alternating current equation for circuits containing only resistance and self-induction is, then,

$$I = \frac{E}{\sqrt{R^2 + 4\pi^2 f^2 L^2}}.$$

The impedance will be higher for high frequencies than for low, which means that in this type of circuit currents of higher frequencies will be cut down more than those of lower frequency. If the frequency becomes zero (that is, if the current is continuous), then the reactance member of the denominator, $4\pi^2 f^2 L^2$, becomes zero and the formula becomes $I = E/R$, which is Ohm's law. On the other hand, no matter what the frequency, if the inductance L becomes zero, the reactance member becomes zero, and the formula again becomes $I = E/R$ and applies to any frequency. Thus Ohm's law in its simple form applies, even in inductive circuits, to continuous currents and in non-inductive circuits to alternating currents of ordinary frequency and voltage.

Applying these principles directly to telephony, we see that ordinarily the presence of an inductance coil, such as that of an electromagnet, in the path over which voice currents must pass, will be deleterious to good speech transmission. The self-induction of the magnet tends not only to reduce the flow of all varying currents but also suppresses the higher overtones more than the lower ones and the fundamental. To express this telephonically, the presence of an inductance in series in the path of voice currents not only *attenuates* the transmitted current but also *distorts* the speech as well. There is an important exception to this general statement which will be referred to later in this chapter.

On the other hand, the connecting of coils having high coefficients of self-induction in shunt or bridge paths across the two sides of a circuit over which voice currents must pass is not so harmful, because the high impedance offered by these electromagnets prevents any large fraction of the voice currents from being shunted from the path it is desired to have them follow.

Again, the property of an inductance coil by which it offers high impedance to high frequency currents and comparatively

low impedance to continuous currents is often made use of in telephone systems in cases where the rapidly fluctuating currents, used in speech transmission, and the continuous currents, used for signaling or other purposes, must follow common paths and then separate each into a path of its own. If a coil of high inductance but low resistance is placed in a circuit common to the two kinds of current, the continuous current will flow freely through it, but the enormous impedance it offers to the voice currents will constitute an effective barrier to their flow and compel them to take another path of such high ohmic resistance as effectually to bar the passage of the continuous current.

2. Reactance Due to Electrostatic Capacity.—Coming now to that reactance which is not due to electromagnetism but to electrostatic capacity, we may again consider a few fundamental principles, this time concerning to electrostatic induction.

Every insulated conductor is capable of receiving a certain charge when subjected to an electromotive force. For instance, if a metallic plate insulated from all surrounding bodies is connected to one terminal of a battery the other terminal of which is grounded, a certain amount of electricity will flow to the plate until its potential is raised to that of the battery terminal. The plate is then said to be *charged*, and the amount of electricity held by it is proportional to its *capacity*. The charge of electricity on the plate would be considered positive or negative, according to whether the positive or negative terminal of the battery, or other charging source, was connected to it.

No charge exists by itself—there is always an opposite charge induced by it upon neighboring bodies. It is known that like charges repel each other, while unlike charges attract; that if an uncharged body be brought near a charged body, an opposite charge will be induced on the side of the uncharged body which is toward the charged body, and that, similarly, a charge of the same sign as that on the charged body will be driven to the opposite side of the uncharged body. If the body which was originally uncharged is connected with the ground, this last charge, namely the one of the same sign as the original, will be driven to the ground, while the charge of opposite sign will still be attracted or "held bound" by the charge on the first body. The second body will, therefore, be charged although it has not been in contact with the first. The action between charges of

electricity taking place through an insulating medium is called *electrostatic induction.*

It is found that when two conductors are placed side by side, but insulated from each other, the capacity of each to hold a charge will be greater than if the other were not present. For the purpose of holding charges in this way the well-known Leyden jar (Fig. 220) has long been in use. It is usually made by coating a glass jar inside and out with a layer of tin foil to within a few inches of the top. The outer coating is generally connected with the ground, while the inner one is connected with a metal rod extending to it through the mouth of the jar. If the inner coating is connected with a source of electromotive force, a current, lasting but an instant, will flow into the coating, putting a charge on it. This charge, which we will say is positive, will attract a nearly equal negative charge to the outer coating, repelling an equal positive charge to the earth, as already described. The amount of charge which the inner coating will receive under these circumstances is very much greater than if the outer coating were not present, and the capacity of the inner coating is, therefore, much higher. Devices consisting of two electrical conductors separated by an insulating medium for the purpose of holding two charges of electricity are called "condensers." The Leyden jar is, therefore, a type of condenser.

In general the capacity of a condenser is increased if the area of the conducting surface is increased, or if the distance between the conductors is diminished, and may be increased or diminished within certain limits by using different kinds of insulating material between the conductors. The medium separating the conductors is called the *dielectric*, and upon it depends to a very great extent the efficiency of the condenser. Several condensers built exactly alike, so far as size of plates and the distance between them are concerned, but using different materials for dielectrics will be found to have different capacities. This difference is due to a property, possessed in varying degrees by different dielectrics, called *inductive capacity.* It is the ability of a dielectric to permit inductive action to take place through its mass and is a quality inherent in the material itself.

For practical purposes the inductive capacity of dry air at atmospheric pressure and at a temperature of 0° C. maybe taken as unity, and the inductive capacities of other dielectrics measured in terms of this unit. When so measured it is called *specific*

inductive capacity. The specific inductive capacity of a given substance is then the *ratio* of the capacity of a condenser whose plates are separated by that substance to the capacity of a condenser exactly like it, except that its plates are separated by a layer of dry air. To illustrate: If a given condenser with a dielectric of plate glass has a capacity four times as great as a similar condenser of exactly the same dimensions but with air instead of glass between its plates, the specific inductive capacity of that particular piece of glass would be 4. Examples of the specific inductive capacities of a number of substances which are sometimes used as dielectrics, are given in Chap. XIV.

To sum up, the capacity of a condenser is the measure of its ability to hold charges of electricity on its plates for a given difference of potential between them. It varies directly as the area of the plates, inversely as the distance between them and directly as the specific inductive capacity of the dielectric separating them.

Specific inductive capacity is often a very important consideration in the choice of material for insulating purposes in telephony. For instance, in the construction of telephone cables it is desirable, as will be shown later, to keep the electrostatic capacity between the two wires of a pair as low as possible. In order to do this the wires are insulated with a very porous dry paper so that the insulation contains a large amount of dry air, with just enough solid substance to keep the wires from touching each other. On the other hand, in the construction of most of the condensers used in telephony, it is desired that the capacity be as great as possible for a given arêa of plates and, therefore, the paper insulation is impregnated with some substance, such as paraffin wax, which has a specific inductive capacity approximately twice that of dry air.

The unit of electrostatic capacity is the *"farad."* A condenser which is capable of holding one *coulomb* of electricity at a pressure of one volt is said to have a capacity of one farad. It will hold twice as much electricity at a pressure of two volts, but its capacity is still one farad. As an analogy, a quart bottle will hold twice as much gas at two pounds pressure as at one, yet the capacity of the bottle remains one quart.

The definition of the farad becomes more significant when it is remembered that the coulomb is the unit of *quantity* of electricity and is the amount represented by a flow of one ampere for one

second. From this it follows that an ampere of current flowing into a condenser of one farad capacity for one second will raise the potential across the terminals of the condenser one volt. A flow of two amperes for one second or of one ampere for two seconds would put two coulombs of electricity into the condenser, and raise the potential two volts, but the capacity would remain at one farad.

Obviously, the amount of electricity a condenser will hold depends on the pressure or voltage as well as on the capacity. Thus a condenser of K farads will hold a charge of K coulombs at one volt and of KE coulombs at E volts; just as a bottle of Q quarts capacity would hold Q units of air at a pressure of one pound and QP units of air at a pressure of P pounds.

A condenser having a capacity of one farad would be of enormous size. The unit is, therefore, too large for ordinary convenience. The condensers most used in telephony have capacities that most conveniently may be expressed in millionths of a farad or in *microfarads*, the prefix *micro* signifying the "one millionth part of." In radio work, where very much smaller capacities are often dealt with, the unit is still further subdivided.

Like inductance, capacity or capacitance must evidently be thought of as a characteristic of the circuit and not as one of the current. This does not mean that the capacity must necessarily be embodied in what we ordinarily call a condenser with flat plates separated by a thin dielectric. Some capacity is inherent in every circuit, even where none is intended or desired. The conductors of the circuit exhibit capacity effects with respect to each other, or with respect to other near-by conductors, such as the earth.

Frequently, these capacities which are incidental to a circuit are inappreciably small and may be ignored, but in some cases there are of great moment. As an example of the latter class, long telephone lines may be cited, particularly where the wires are brought close together, as in cables. The two wires of the line act as the two plates of a condenser, and the air or paper insulation between them as the dielectric. It will be shown that this capacity inherent in telephone lines exerts a deleterious effect on voice transmission. One of the outstanding achievements of the scientist and the telephone engineer has been the over coming of trouble due to this cause.

As in the case of electromagnetic inductance, we may profitably consider what happens in a circuit containing capacity: first, when the electromotive force is continuous and, second, when it is alternating.

A circuit containing a constant source of electromotive force (battery) is shown in Fig. 98. Here all the resistance of the circuit is supposed to be concentrated in R and all the capacitance in the condenser C. The insulation of the dielectric between the condenser plates is assumed to be perfect, so that no current can flow through it from one plate to the other. Notwithstanding this complete conductive break in the circuit, a momentary current will flow in its conductors when the battery electromotive force is applied, as by closing Key No. 1, and another will flow in the opposite direction when the battery electromotive force is removed, as by closing Key No. 2. These currents are, of course, the movements of electricity on to and off of the condenser plates while they are charging and discharging.

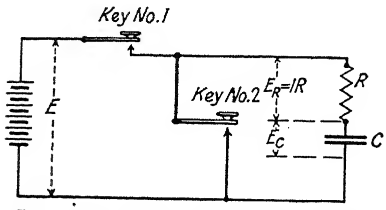


FIG. 98.—Circuit containing constant electromotive force, resistance and capacitance.

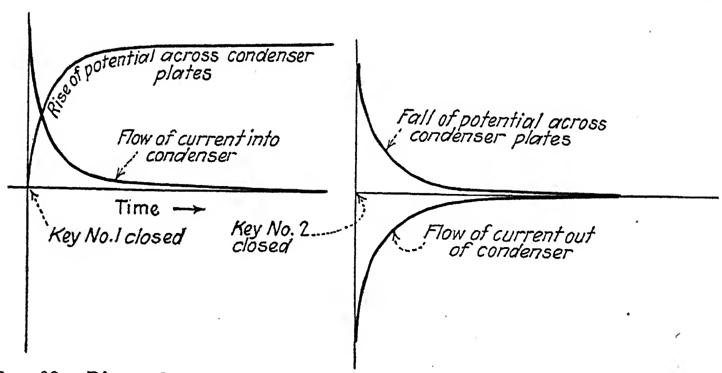


FIG. 99.—Rise and fall of current in circuit having resistance and capacitance.

A graphical illustration of these momentary currents in such a circuit and of the rise and fall of the potential across the condenser plates following the application and removal of the battery electromotive force is shown in Fig. 99. The curves at the left

relate to the period of charge, those at the right to the period of discharge.

Taking first the period of charge: If there were merely a gap in the circuit, instead of the condenser, with no capacity in the conductors leading to it (not practically possible), no current would flow upon the closing of Key No. 1, and the potential across the gap would *instantly* become equal to that across the battery terminals. On the other hand, if the condenser were short-circuited, a current, $I = E/R$, would flow the instant Key No. 1 was closed and continue to flow as long as it remained closed. With the condenser in the circuit, however, the current *starts* to flow at that same rate, as determined only by the electromotive force and the resistance; for at that instant the condenser, having no charge, has no counter electromotive force with which to oppose the flow. With the flow of current during the next succeeding instants, electricity gradually[1] accumulates on the condenser plates, and as its quantity increases the potential between the condenser plates gradually increases, being, of course, in such direction as to oppose the applied or "impressed" electromotive force of the battery. Finally, when the condenser can hold no more at the applied voltage, the current ceases, because the counter electromotive force of the condenser has now risen until it is equal to that of the battery. The sum of the electromotive forces acting in the circuit has become zero. Equilibrium again exists, but energy conditions are different from those at the beginning of charge. Two kinds of work have been done: that used up in heating the resistance and that used in building up the electrostatic field across the dielectric. The former work is lost in heat; the latter is merely a storage of energy, like the energy of a compressed spring, to be recovered when the condenser is allowed to discharge.

This electrostatic field built up in the charging of the condenser is really a state of stress across the dielectric. If the stress becomes too severe, as by the application of too high an applied voltage, it may actually cause a disruption of the material of the dielectric, a breakdown of the insulation and a resulting flow of current across the break until the potential is reduced.

[1] The word "gradually" is used advisedly in spite of the fact that the time involved in the charging of the condenser may be so small as to be most conveniently measured in millionths of a second, or even smaller units of time.

It takes work to create such an electrostatic field and we may think of it in much the same way as that it takes work to pump up an automobile tire. In the case of the condenser electricity is forced into it, creating an electrical pressure which is resisted by the strength of its dielectric; in the case of the tire air is forced into it, creating a pressure which is resisted by the strength of its fabric. In both cases energy is stored, and in both cases this energy will be given back if a path is afforded through which the pressure may be relieved.

On discharge the conditions are reversed. When Key No. 2 of Fig. 98 is depressed, current flows out of the condenser through the short-circuit path provided, and the plate potential gradually drops as the charge is lost. The successive instantaneous values of current and potential are indicated by the curves at the right of Fig. 99. The current curve now lies below the horizontal axis, indicating that the flow is in the opposite direction from that during charge and, therefore, negative. The difference of potential across the condenser plates during discharge is of the same sign as that during charge and as before, is indicated as positive.

At the instant of starting the discharge the current is at its maximum negative value, since the battery electromotive force has been entirely withdrawn, and the full potential across the condenser plates still exists. As electricity continues to flow out of the condenser, however, the potential across its plates decreases and with it the current. Equilibrium is finally again reached when the potential difference and the current have reached zero values. The electrostatic field has now been destroyed, and conditions are again the same as before the charge began. In restoring them two kinds of work have been involved, that used up in heating, which is lost, and that which was stored in the electrostatic field which, like that of a released spring, is returned to the circuit.

Obviously, since the resistance in the circuit limits the rate at which electricity can flow under the influence of a given electromotive force, the time required to charge or to discharge the condenser will be proportional to the resistance of the circuit. Also the time required will be proportional to the capacity of the condenser. In the same way, the time required to pump up or to empty an automobile tire will be greater if there is a restriction in the supply pipe and will be greater for a large tire than

for a small one. For a given circuit, therefore, the time required to bring a condenser to a given state of charge or discharge is proportional to the product of the resistance and the capacity. The product RC, where R is in ohms and C is in farads, is called the *time constant* of the circuit.

A fundamental distinction may be noted here between the action of electrostatic capacity and of self-induction in a circuit. In the case of a circuit containing only resistance and self-induction (Fig. 91 and 92) it was found that changes in current always lagged behind changes in electromotive force. In the present case, however, in a circuit containing only resistance and capacitance (Figs. 98 and 99) the rise and fall of the potential across the condenser terminals lags behind the changes in current. In other words, the *current leads the electromotive force.* This is best seen from the curves of Fig. 99. At the left, during the charging period, the potential of the condenser terminals does not rise to its final value until a time after the current has had its maximum value. At the right, or discharge, the potential does not reach its final (zero) value until after the current has had its maximum value.

So far, only a continuous electromotive force has been considered in connection with condenser circuits. Before passing to the consideration of alternating electromotive forces, it is well to point out that although the impressed voltage in Figs. 98 and 99 was constant, the voltage actually applied to the condenser terminals was variable during the periods of charging and discharging, as shown by the potential curves of Fig. 99. The reason for this was that the drop IR through the resistance R varied with the current.

In order to discuss the relationship between the condenser current and a sinusoidal electromotive force applied directly to its terminals, consider Fig. 100, where the assumption is made (only approximately possible) that the circuit has only capacitance. Here, there being no resistance, the impressed electromotive force is applied directly to the condenser plates. The condenser plates will follow the potentials thus applied, and the condenser will be alternately charged and discharged accordingly. The movement of electricity into and out of the condenser will constitute the alternating current flowing in the circuit. There can be no other current flowing because the condenser dielectric is assumed to be a perfect insulator. It is obvious that as the

potential applied across the plates is raised, current will flow into the condenser, because it holds more electricity at high potentials than at low. When the applied voltage ceases to change, no current flows one way or the other, and, when the applied voltage decreases, current flows out, because so much electricity cannot be held at lower potentials.

We see then, that the current at any instant does not depend on the electromotive force across the condenser terminals at that instant, but rather upon whether the electromotive force is changing or not. An important law may here be observed with respect to the flow of alternating currents in condenser circuits. It is not

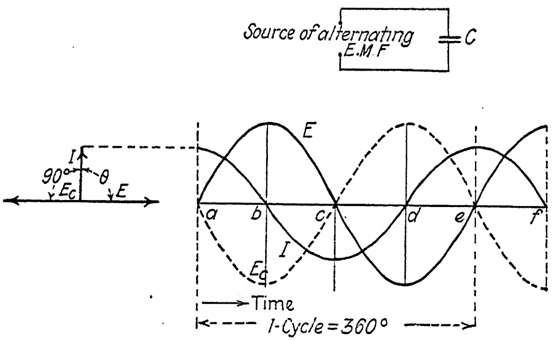


FIG. 100.—Current and electromotive force waves in circuit having capacitance only.

the magnitude of the potential applied to the condenser plates at any instant that determines the current flow, but *the rate at which the applied potential is changing.* If the potential stops changing, current will cease, no matter what the potential value at that instant. Furthermore, the direction of the current will, obviously, depend on whether the applied electromotive force is increasing or decreasing.

Referring to the two full-line curves E and I of Fig. 100 and considering, as usual, that positive values are shown above the axis and negative below, it is evident that at the instant indicated by a on the horizontal axis the electromotive force at the condenser terminals is changing most rapidly, and, therefore, although the potential is zero at this time, the current will be at its maximum. This current is taken as positive in direction because on a rising potential it is flowing into the condenser. At b the electromotive force, though at its maximum, has ceased to change, and, therefore, the current has dropped to zero.

From b to d the applied electromotive force decreases from its maximum positive to its maximum negative value. Under decreasing potential, current flows out of the condenser and is negative. It reaches its maximum negative value at c, as the electromotive force is passing through zero, for at that point the electromotive force is falling most rapidly. From d to f the applied electromotive force is again rising. Current flows into the condenser during this period and is therefore positive in direction.

An inspection of these curves of Fig. 100 will show that, as already has been pointed out, the current leads the electromotive force. That is, it reaches any given phase of its cycle before the electromotive force reaches its corresponding phase. Furthermore, it is clear that the phase difference is just 90 degrees, since the current passes through a given point in its cycle just a quarter period in advance of the time when the electromotive force passes through the corresponding point in its cycle.

As a condition of no resistance in the conductors of the circuit was assumed, there is no resistance drop IR opposing the impressed electromotive force. The only opposition to the impressed electromotive force is the counter electromotive force of the condenser, which at every instant must be just equal and opposite to the impressed electromotive force. This counter electromotive force, or condenser electromotive force, E_c will thus be represented by the dotted curve of Fig. 100, obviously having phase difference of a half period or 180 degrees with the impressed electromotive force.

The three curves E_c, I and E of Fig. 100 may be considered as being generated by the revolution of the three vectors E_c, I and E at the left of the diagram. Clearly, the condenser electromotive force must always lead the current I by 90 degrees whether there is resistance in the circuit or not, for it is wholly dependent on the current that is flowing. The angle between vectors E_c and I must, therefore, always be a right angle. The angle θ, however (called the "angle of lead"), by which the current I leads the applied electromotive force, can never be as great as a right angle except under the limiting condition of a resistanceless circuit such as we have assumed.

Under the no-resistance condition assumed, we may show that the current would be truly wattless. At any instant the power would be the product of the electromotive force and current

values at that instant, or $P = EI$. At points a, b, c, d and e where either curve crosses the horizontal axis, the power would be zero because, at those instants, one factor or the other, E or I, of the power equation would be zero. At all other times the instantaneous values of the power would be finite, and positive or negative according to whether the current and electromotive force values were of the same or of opposite signs. We may draw the power curve as in Fig. 101, where curves E and I are the same as in Fig. 100. At a, where current is maximum and electromotive force zero, the power is zero. At b, where the current is zero and the electromotive force a maximum, the power is again zero. But at all points between a and b the power is finite and positive because both E and I are finite and positive. From b to c the power will be negative because E is positive and I negative. Thus, in the circuit of Fig. 100, when the product EI is positive, power flows from the source into the condenser, and when the product is negative, the power flows back again. Since the areas of the power curve above and below the zero line are equal, the net exchange of power is zero. The power is never zero except at four instants during each cycle. The successive quarter cycles are alternately positive and negative, but the average power for the cycle is zero. The net result is what we term a "wattless current."

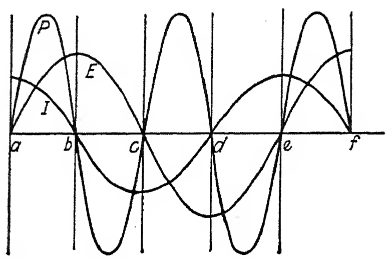


FIG. 101.—Power curve in circuit having capacitance only.

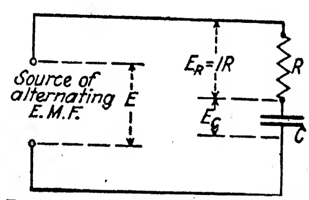


FIG. 102.—Circuit containing resistance and capacitance.

That this current is wattless may also be seen from the power equation $P = EI \cos \theta$, already mentioned as applicable to sinusoidal currents not in phase with their impressed electromotive forces, for, as will be remembered, the cosine of an angle of 90 degree is zero. Here, as in the case of a circuit containing only self-induction, the current surges back and forth without resistance and, therefore, with no consumption of energy.

In Fig. 102 a simple circuit containing resistance and capacitance is shown. The impressed electromotive force is no longer applied directly at the condenser terminals, it being always modified by the varying drop *IR* through the resistance *R*. We know, however, that the condenser electromotive force, that is, the counter electromotive force offered by the condenser, will be 90 degrees in advance of the current actually flowing. There are, then, two sinusoidal electromotive forces acting to drive current through the resistance *R*, namely, the impressed electromotive force *E* and the condenser electromotive force E_c.

The relationships of these electromotive forces to each other

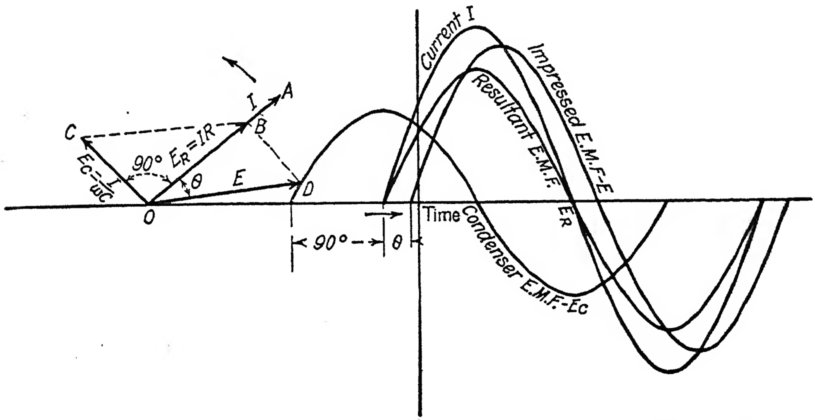


FIG. 103.—Current and electromotive force waves for circuit having resistance and capacitance.

and to the current may be analyzed with the aid of the curves and vector diagram of Fig. 103. Let the current flowing as a result of these two electromotive forces be indicated by curve *I*, which may be conceived as being generated by the rotation of the vector *OA* about the point *O*. The length of *OA* will represent the maximum instantaneous value of the current.

The electromotive force that is driving this current through the resistance *R* will, as just stated, be the resultant of the impressed and the condenser electromotive forces. It will, of course, be in phase with the current and its value, according to Ohm's law will be *IR*. We may, therefore, lay off vector *OB* equal to *IR* and coincident in direction with *OA*. The curve of this resultant electromotive force E_R may be conceived as being

generated by the rotation of this vector OB. It reaches its zero and maximum values simultaneously with the corresponding values of its current curve.

Curve E_c of Fig. 103, representing the condenser electromotive force, leads the current by 90 degrees. On the vector diagram it is represented by the vector OC at right angles to, and leading, the current vector. Its length will represent the maximum value of the condenser electromotive force. The method of determining this maximum value will be dealt with presently.

We now have on the vector diagram the resultant electromotive force E_R and one of its components E_c. To find the other component, the impressed electromotive force E, we have only to complete the parallelogram of forces by drawing OD parallel to CB and BD parallel to OC. Clearly, then, OD represents the impressed electromotive force both in maximum value and in direction. Its corresponding curve E shows the instantaneous values and phase relationship of the impressed electromotive force.

It is apparent, either from the curves or from the vector diagram of Fig. 103, that the current flowing in a circuit containing only resistance and capacity will lead the impressed electromotive force by an angular amount which may be measured along the horizontal axis of the curves of instantaneous values or, more directly, by the angle θ between the vectors. Obviously, this angle must always be less than 90 degrees, since the resistance drop $E_R = IR$ must always be present in some degree.

The maximum value of the condenser electromotive force E_c, needed to determine the length of the vector OC in Fig. 103, may be arrived at from the following consideration:

The amount of electricity Q that a condenser will take as a charge is proportional to the capacity C and the voltage E applied to its terminals. $Q = CE$ when Q is expressed in coulombs, E in volts and C in farads. Also, the amount of electricity in a condenser charge is proportional to the product of the current flowing into the condenser and the time of the flow. $Q = IT$ when I is in amperes, T in seconds and Q in coulombs. It follows, then, that $IT = CE$ or $I = CE/T$. If the time is one second, then $I = CE$. This is true for a *steady* flow I lasting one second; or, if the current is varying, for an *average* flow I lasting one second.

In the case under consideration the electromotive force at the condenser terminal is the varying (sinusoidal) electromotive force E_c of Fig. 103. The condenser, therefore, receives a charge CE_c twice and discharges it twice during each cycle. The total quantity of electricity flowing in a cycle is, therefore, $4CE_c$, and if the period of the cycle is one second, then $I = 4CE_c$. If there are f cycles per second (frequency $= f$), then

$$I = 4fCE_c.$$

This simply means that an amount of electricity equal to $4fCE_c$ is flowing in the circuit each second disregarding direction or manner of variation. It is, therefore, the ordinary *average* value of the current flowing.

Since the ratio of the average value to the maximum value in harmonic variation is always as 1 to $\pi/2$, it follows that when the variation of current is harmonic the maximum instantaneous value of the current will be

$$I = 4fCE_c \times \frac{\pi}{2} = 2\pi fCE_c.$$

This also may be written

$$I = \omega CE_c.$$

where ω is the angular velocity and equal to $2\pi f$. In this equation, as just derived, I and E_c are maximum instantaneous values, but since the ratio of maximum values to effective (root-mean-square) values is the same for both current and electromotive force, it is equally true of effective values. This equation amounts merely to a statement of the obvious fact that the current flowing into and out of a condenser is proportional to the electromotive force applied to the condenser terminals, to the capacity of the condenser and to the rapidity with which the charges of electricity are accumulated and rejected.

The last equation may also be written

$$E_c = \frac{I}{\omega C},$$

and this is the value needed to give the length of the vector OC already referred to in connection with Fig. 103. For this purpose E_c and I are maximum instantaneous values.

As in the study of self-induction, we may always indicate the relationship, both quantitative and directional, of the

electromotive forces acting in a circuit containing resistance and capacity by means of a right triangle taken from such a parallelogram of forces as that of Fig. 103. Such a triangle of electromotive forces stripped of details is shown in Fig. 104.

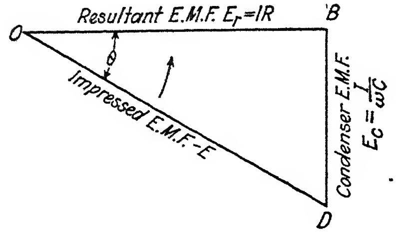


FIG. 104.—Electromotive force triangle for circuit having resistance and capacitance.

The three sides represent, by their lengths, the values of the respective electromotive forces, and, by their directions, the directions in which these electromotive forces are acting. It always will be a right triangle because the condenser electromotive force is always at right angles to the current, which is in phase with the resultant electromotive force. Also, the angle between the impressed and the resultant electromotive forces will always represent the "angle of advance" by which the current leads the impressed electromotive force.

From the right triangle DOB of Figs. 103 or 104 we see that

$$\tan\theta = \frac{DB}{OB} = \frac{I/\omega C}{IR} = \frac{1}{\omega CR} = \frac{1}{2\pi fCR}.$$

This means that in a circuit containing only resistance and capacity in series, the current will lead the impressed sinusoidal electromotive force by an angle whose tangent is $1/\omega CR$ or $1/2\pi fCR$. The tangent of the angle of lead thus varies inversely as the frequency, the capacity and the resistance and, therefore, inversely as the product of the three.

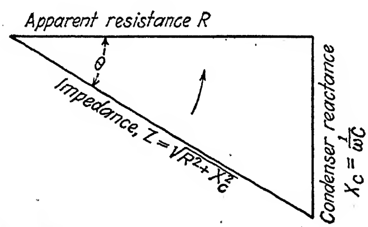


FIG. 105.—Impedance triangle for circuit having resistance and capacitance.

An exactly similar triangle (Fig. 105) is obtained by dividing each side of the electromotive force triangle (Fig. 104) by I. As in the case of Fig. 97, this may be called the "impedance triangle," but here the impedance is due to resistance and capacity instead of resistance and self-induction. The three sides of this triangle show, respectively, the opposition to the flow of current offered by the apparent resistance R, the condenser

reactance X_c and their resultant the impedance Z; also the relative direction in which each is exerted.

The condenser reactance $X_c = 1/\omega C = 1/2\pi fC$ is the choking effect of the condenser. It acts always at right angles to the resistance, vectorially speaking, and varies inversely as the frequency, which is a characteristic of the current, and inversely as the capacity, which is a characteristic of the circuit. In other words, the higher the frequency or the greater the capacity the less the opposition offered by the condenser to the flow of alternating currents. Also, it is to be noted that the condenser reactance is independent of the resistance.

The total impedance Z is, according to the law of the right triangle, equal to the square root of the sum of the squares of the resistance and the condenser reactance. Thus we have:

$$\text{Impedance } Z = \sqrt{R^2 + X_c^2} = \sqrt{R^2 + \frac{1}{\omega^2 C^2}} = \sqrt{R^2 + \frac{1}{4\pi^2 f^2 C^2}}.$$

The current equation for circuits containing only resistance and capacity is, then,

$$I = \frac{E}{\sqrt{R^2 + \frac{1}{4\pi^2 f^2 C^2}}}.$$

An examination of this equation will show the effect of frequency on current flow in circuits containing resistance and capacity. Obviously, the higher the frequency the lower will be the impedance. If the frequency is reduced to zero (direct current), the condenser reactance and, therefore, the impedance become infinite. The current is therefore zero. On the other hand, if the frequency becomes infinite, the second term of the impedance radical becomes zero and the current equation becomes merely $I = E/R$, the current being limited only by the resistance. In general, currents of high frequency will encounter less impedance than those of lower. This statement is to be contrasted with that concerning the reactance due to self-induction, where it was shown that exactly the opposite condition was found in circuits containing resistance and self-induction.

Considering the very general application of these principles of the condenser to the practice of telephony, a condenser of ordinary size may be placed in series in the path over which voice currents must pass without offering serious opposition to their flow. This is because the voice currents are of suffi-

ciently high frequency to make the second term of the impedance radical $\sqrt{R^2 + \frac{1}{4\pi^2 f^2 C^2}}$ very small and, consequently, of little importance in comparison to the resistance. Of course there is some attenuation and some distortion offered to voice currents by a series condenser. There is some attenuation because we are dealing with finite frequencies, and, therefore, the condenser reactance cannot quite disappear. The distortion results from the fact that the voice currents are comprised of a large number of component waves of different frequencies, and there is a tendency to suppress the lower frequencies more than the higher. These effects under ordinary circumstances and for condensers of ordinary size, however, are relatively slight.

The connection of condensers across or in shunt relation to a path carrying voice currents involves quite a different situation. Bridged condensers present effective shunt paths for the voice currents, thus greatly reducing the amplitudes of the currents arriving at the distant end of the circuit. Obviously, the larger the capacity of the condenser and the higher the frequencies the greater this shunting effect will be. Ordinarily, therefore, we do not willfully bridge condensers across a telephone line. Unavoidably, however, the telephone engineer is confronted with just this situation of bridged capacity across every telephone line he builds. This is because in each section of the line the two conductors with their interposed insulation form a condenser, just as any other two conductors insulated from each other would do. The condition is that suggested in Fig. 106. Each short length of line conductors form, in effect, the plates of a condenser and the insulation between them the dielectric. The result is "distributed capacity." For short lines the shunting effect of the condensive action of the line is not serious; for long lines it is.

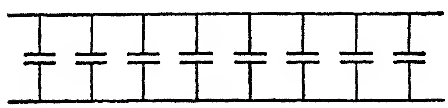


FIG. 106.—Suggesting the capacity effect of line conductors.

On the other hand, the condenser forms a valuable means for keeping the direct or low-frequency currents used for signaling separate from the comparatively high-frequency currents used in talking, in cases where the two kinds of currents

must follow common paths and then separate into paths of their own. It bars the direct currents, offers high impedance to low-frequency signaling currents and provides a ready path for the higher frequency currents used in voice transmission.

Inductance and Capacitance Compared.—It must have become apparent from the earlier portions of this chapter that in many respects self-induction and capacity behave in exactly opposite ways in their influence on the flow of current. The condenser permits the flow of alternating currents, the inductance tends to bar them; the condenser is a barrier to direct currents, the inductance offers little opposition to their flow; condensers offer less opposition to higher frequencies, inductances more; and, lastly, condenser currents lead in phase while inductance currents lag.

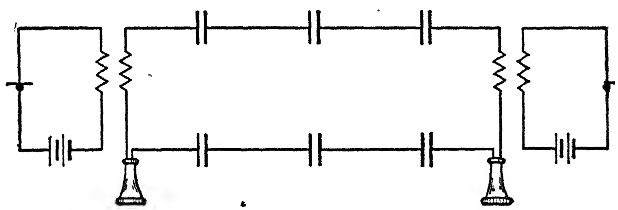


Fig. 107.—Telephone line with condensers in series.

The simple illustrations of Figs. 107 to 110 show these different characteristics of condensers and inductance coils with respect to the transmission of voice currents. Conceive the two telephones to be connected by a line of two continuous conductors so short and well insulated that its resistance, leakage, inductance and shunt capacity are practically negligible. If a number of condensers of, say, one microfarad each be introduced in series in the line as shown in Fig. 107, little or no noticeable effect is produced on the transmission between the two telephones, although the conductivity of the line is practically zero.

If the condensers be connected in shunt across the line, as in Fig. 108, the listener at the receiving end hears greatly reduced transmission, while the volume in the speaker's receiver is considerably augmented. The line, in effect, has been short-circuited with respect to voice currents, although continuous currents could pass over it quite as effectively as before. This presents an exaggerated case of the distributed capacity of which every line is inherently possessed in some degree.

In Fig. 109 instead of serially connected condensers we have serially connected inductance coils. If the inductance of the coils is high, even though their resistance is very low, there is great interference with the transmission.[1] The coils not only cut down the volume of sound heard in each of the receivers (attenuation) but they also suppress the higher harmonics to a greater extent than the lower, thus producing distortion

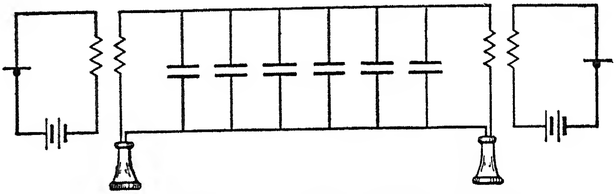


Fig. 108.—Telephone line with bridged condensers.

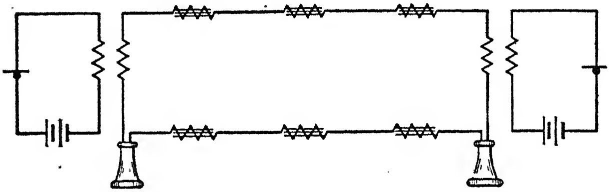


Fig. 109.—Telephone line with inductance coils in series.

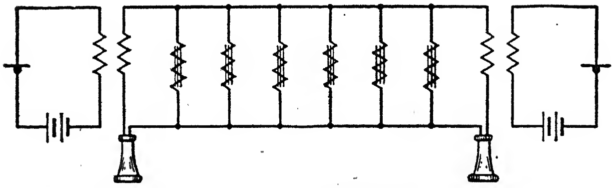


Fig. 110.—Telephone line with bridged inductance coils.

which further interferes with intelligibility. It was in this manner (series) that the coils of the signal bells at the stations on party lines were connected in the talking circuit in the early days of telephony until Mr. J. J. Carty showed that the proper way to connect them was in bridge relation *across the line.*

Such a bridging connection of the inductance coils is shown in Fig. 110. Connecting condensers in this way (Fig. 108)

[1] Nevertheless, as will be shown, specially designed inductance or impedance coils are serially connected at stated intervals in the line to effect a marked improvement in transmission. This is known as "loading" and is to be referred to later.

tends to ruin the transmission; but such coils, even when of low resistance, if they are properly designed to have high inductance, may be connected across the two sides of the line without materially affecting the transmission. Such a line would be most inefficient for transmitting direct currents but may be very efficient for voice currents, because the high reactance of the coils constitutes a practical bar to the passage of rapidly varying currents through the shunt paths.

Combined Effects of Inductance and Capacity.—The opposite effects of inductance and capacity with respect to circuits carrying alternating currents may now be given more extended attention.

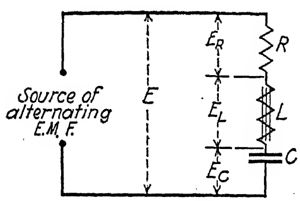


FIG. 111.—Circuit containing resistance, inductance and capacitance.

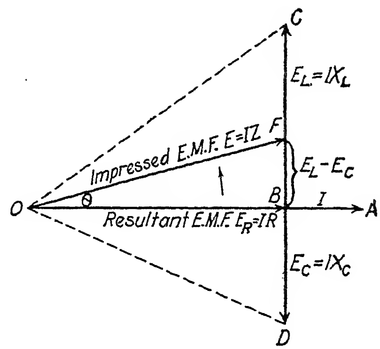


FIG. 112.—Vector diagram of electromotive forces acting in circuit of Fig. 111.

Consider such a circuit as that of Fig. **111** where the three possible components of impedance—resistance, inductance and capacity—are connected in series.

This being a series circuit the current in all parts of it will be the same and will, of course, be in phase with the electromotive force E_R across the resistance. As before, we may consider this electromotive force E_R as the resultant of all the other electromotive forces acting in the circuit. These other electromotive forces, E, that impressed by the source, E_L, that of the inductance, and E_c, that of the condenser, will not in general be in phase with the current or with each other. All must therefore be treated vectorially, as in Fig. 112.

The current I may conveniently be represented by the horizontal line OA. The electromotive force across the resistance, equal to the product IR, is measured by the vector OB coincident

in direction with OA. The electromotive force of self-induction E_L is represented by the vector BC laid off at right angles to the current and leading (*i.e.*, current lags 90 degrees behind it). Similarly, the electromotive force of the condenser E_c is represented by the vector BD, at right angles to the current and lagging (*i.e.*, current leads it by 90 degrees).

The impressed electromotive force would be represented by the dotted vector OC if the inductance only were present, and by the dotted vector OD if the capacitance only were present. With both present and acting in opposite directions, it is evident that its vector will lie between these two extremes. Since E_L and E_c are 180 degrees apart in phase (perfect condenser and pure inductance assumed), they act oppositely in the same straight line, and their combined effect is equal to their arithmetical difference. This difference $E_L - E_c$ representing the combined effect of the electromotive forces of inductance and capacity is represented on the diagram by vector BF, and the triangle of electromotive forces is completed by vector OF representing the impressed electromotive force. Under the conditions assumed in the diagram the inductance predominated over the capacity, so that E_L was greater than E_c. Their difference, therefore, acted in the direction of E_L, that is, upward from point B. This places the impressed electromotive force ahead of the current in phase, and the angle θ indicates the angle of current lag. Had the condition been reversed, with E_c larger than E_L, their combined effect $E_L - E_c$ would have acted in the direction of E_c, that is, downward from B. This would place the impressed electromotive force behind the current in phase, and the angle θ would have been the angle of current lead.

From the triangle OBF of Fig. 112 we obtain, directly,

$$E = IZ = \sqrt{I^2R^2 + (IX_L - IX_c)^2}.$$

Dividing by I, the impedance Z is determined, as

$$Z = \sqrt{R^2 + (X_L - X_c)^2}.$$

Substituting the values of X_L and X_C in terms of angular velocity, frequency, inductance and capacitance, we have

$$Z = \sqrt{R^2 + \left(2\pi f L - \frac{1}{2\pi f C}\right)^2}.$$

The current, therefore, is

$$I = \frac{E}{Z} = \frac{E}{\sqrt{R^2 + \left(2\pi fL - \frac{1}{2\pi fC}\right)^2}}.$$

This is Ohm's law for sinusoidal currents in circuits containing resistance, inductance and capacitance in series. Here I and E may be either the maximum or the effective values of current and electromotive force, respectively, while f is the frequency in cycles per second. R, L and C are expressed in ohms, henries and farads, respectively.

Electrical Resonance.—We have seen that in a circuit containing resistance, inductance and capacity, the current lags or leads according to whether the electromotive force of inductance or of capacity predominates. This leads to the interesting conclusion that if the electromotive forces of inductance and capacity are just equal, they will exactly annul each other, since they are always in opposition, and the current will neither lag nor lead. The current will, therefore, be in phase with the impressed electromotive force, even though the balancing electromotive forces of inductance and capacitance each be large. This is quite clear from the diagram of Fig. 112, for if $E_L = E_C$, their difference will be zero. Point F will, therefore, fall on the current vector, making the vector of impressed electromotive force OF coincident with the current vector OA, the angle θ becoming zero.

When $E_L = E_C$, then also $X_L = X_C$. That is, the inductive reactance will balance the condensive reactance. The impedance

$$Z = \sqrt{R^2 + (X_L - X_C)^2}$$

will then have is minimum possible value, equal to the ohmic resistance R, for the quantity in the parenthesis becomes zero, and the impedance equation becomes merely

$$Z = R.$$

Likewise, the current will have its maximum possible value

$$I = \frac{E}{Z} = \frac{E}{R}.$$

In other words, when the inductive reactance just balances the condensive reactance, their effects are eliminated and the resistance is the only obstruction to the flow of current.

Also this will be the condition for maximum power in the case where E remains constant and X_L and X_C are varied. The power equation

$$P = EI \cos \theta$$

becomes merely

$$P = EI,$$

the same as for continuous currents. This is true because here the angle θ is zero, and the cosine of zero is 1.

A little consideration will show that since the inductive reactance $X_L = 2\pi fL$ increases with the frequency and the condensive reactance $X_C = 1/2\pi fC$ decreases with the frequency, the foregoing balance between X_L and X_C can exist under any given set of circuit conditions for one frequency only. Let us see what the relationship must be between the frequency and the inductance and the capacity of a circuit to bring about this minimum impedance to the flow of current.

The condition for minimum impedance is that $X_L = X_C$. Substituting the values of these reactances in terms of frequency, inductance and capacitance already found, we have

$$2\pi fL = \frac{1}{2\pi fC}.$$

Solving this for f,

$$f = \frac{1}{2\pi\sqrt{LC}}.$$

Under this condition a state of *electrical resonance* is said to exist. The circuit is said to be resonant to that particular frequency, just as a piano string is resonant to a certain frequency of air vibrations.

It will be noted from the equation just derived that the frequency at which a circuit will be resonant is, at least by this simple theory, independent of the resistance, the impressed voltage and the current, since neither R, E nor I enters into its terms. Practically, this is not always true, because, as has been shown, the inductance and, in smaller degree, the capacity are in some cases subject to some variation for different current intensities. Also, it is evident that the resonant frequency will vary inversely as the square root of the *product* of L and C. For a given frequency, therefore, either L or C may have any

value, provided the corresponding value of C or L is such as to give the required product.

An interesting, and sometimes dangerous, aspect of electrical resonance is that the electromotive force across the inductance, and the equal and opposing one across the condenser, may greatly exceed the impressed electromotive force. This and other facts regarding electrical resonance may be illustrated by considering a specific circuit of assumed characteristics. Such a circuit, indicated in Fig. 113, is assumed to

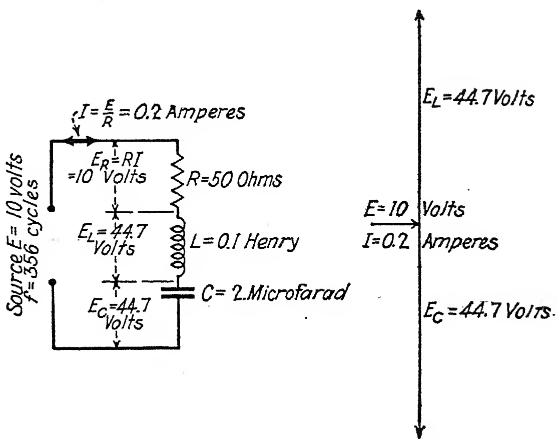


FIG. 113.—Resonant circuit and vector diagram at resonant frequency.

have a resistance of 50 ohms, an inductance of 0.1 henry, and a capacitance of 2 microfarads. These figures are assumed to represent the entire resistance, inductance and capacitance of the circuit. The impressed electromotive force is taken as 10 volts.

We may inquire first as to the frequency at which this circuit will be resonant. It is only necessary to substitute these values of L and C in the equation last discussed. We have then:

$$f = \frac{1}{2\pi\sqrt{LC}} = \frac{1}{2 \times 3.14 \times \sqrt{0.1 \times 0.000002}} = 356 \text{ cycles per second.}$$

At this resonant frequency the current will depend only on the electromotive force and the resistance, and will be

$$I = \frac{E}{R} = \frac{10}{50} = 0.2 \text{ ampere.}$$

The inductive electromotive force at resonance will be

$$E_L = 2\pi f L I = 6.28 \times 356 \times 0.1 \times 0.2 = 44.7 \text{ volts.}$$

Likewise, the capacitive electromotive force will be

$$E_C = \frac{1}{2\pi f C} = \frac{0.2}{6.28 \times 356 \times 0.000002} = 44.7 \text{ volts.}$$

Thus, although the impressed electromotive force is only 10 volts, the electromotive force across the inductance and the condenser are each nearly 45 volts, or nearly four and a half times as great. In some cases resonant circuits may develop electromotive forces several hundred times as great as the impressed, so great in fact as to puncture the condenser dielectrics or do other damage.

The vector diagram representing the resonant condition of the circuit just discussed is shown at the right of Fig. 113. In this, which is drawn to scale for the three electromotive forces acting, the short horizontal arrow represents the impressed electromotive force E. The current I is in phase with this. The electromotive force of self-induction E_L is represented by an arrow at right angles to the current and leading. Similarly, the electromotive force across the condenser E_C is represented by an arrow at right angles to the current and lagging.

It is sometimes a little difficult to gain a physical conception of just what happens in a circuit under such a set of conditions as this, where a comparatively small periodic electromotive force brings into play very much larger forces which just equal and oppose each other. The phenomenon is in some respects analogous to the action of a pendulum. If a very small force is periodically applied to the pendulum at exactly the proper intervals, the pendulum bob, even though of large mass, will be caused to vibrate with relatively wide amplitude. The force which is urging the pendulum to return at either end of its beat is very much greater than the small force which is periodically applied to it. If the small periodic forces are not applied at exactly the intervals demanded by the natural rate of vibration of the pendulum, comparatively little motion will result. In other words, the applied periodic forces must be in resonance with the natural rate of vibration of the pendulum. So it is with a resonant circuit. It has a natural frequency at which the response in current flow will be greatest. If periodic electromotive forces are applied to it at any other frequency, the

current flow will be less; and, in general, the farther removed the applied frequency is from the resonant frequency, the smaller will be the current flow.

We may test this by applying the current equation

$$I = \frac{E}{\sqrt{R^2 + \left(2\pi fL - \frac{1}{2\pi fC}\right)^2}},$$

already derived, for various frequencies at, above and below the resonant frequency. For instance:

At the resonant frequency of 356 cycles,

$$I = \frac{10}{\sqrt{50^2 + \left(6.28 \times 356 \times 0.1 - \frac{1}{6.28 \times 356 \times 2^{-6}}\right)^2}} = 0.2 \text{ ampere},$$

which is in accordance with the value derived from the simpler equation

$$I = \frac{E}{R}.$$

If, however, the frequency is 300, then

$$I = \frac{10}{\sqrt{50^2 + \left(6.28 \times 300 \times 0.1 - \frac{1}{6.28 \times 300 \times 2^{-6}}\right)^2}} = 0.11 \text{ ampere}.$$

A drop in frequency from 356 to 300 cycles per second has thus caused a drop in current from 0.20 to 0.11 ampere for the same applied voltage and in identically the same circuit.

In a similar way the current in this circuit for various frequencies from 100 up to 600 cycles per second have been calculated and are plotted graphically in Fig. 114. The marked maximum in current flow at just 356 cycles is to be noted and also the rapid diminution of current on each side of this natural period as the frequency falls below or rises above this critical value.

The practical bearing of this on telephone transmission is obvious. We have seen that for really high-grade transmission of speech and music a band of frequencies from about 100 to 6,000 must be transmitted, and, as far as possible, all with equal facility. Such a circuit as that of Fig. 113 would pass freely

a narrow band of frequencies at and on each side of 356 cycles but would tend to suppress all others. The higher frequencies

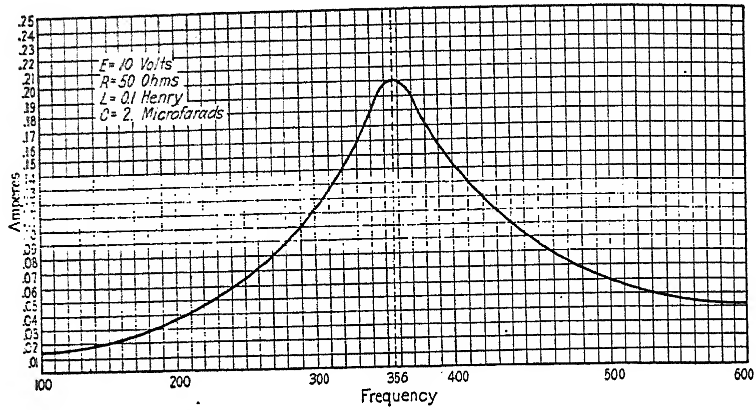


FIG. 114.—Current in circuit of Fig. 113 at different frequencies.

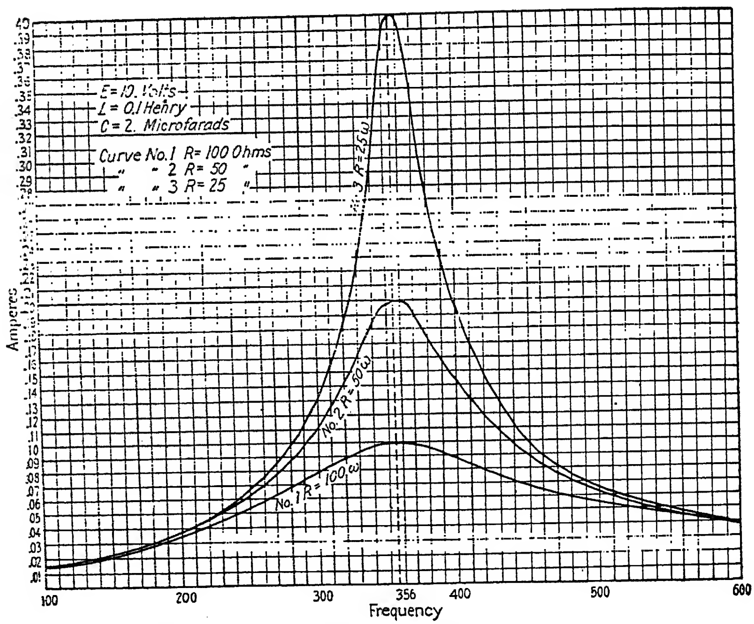


FIG. 115.—Effect of resistance on resonance curve.

from 600 up, vitally important for intelligibility, would evidently be practically barred. The circuit would, however,

be very efficient for transmitting a single tone corresponding approximately to the fundamental of the note f' in the middle octave of the piano. The harmonics of this tone would be largely eliminated and the true quality of the tone completely lost.

Changing the resistance of such a circuit, while not altering the frequency at which it will be resonant, will produce a decided change on the form of the curve showing the relation between current flow and frequency. This is illustrated in Fig. 115 where resistances of 25, 50 and 100 ohms have been assumed without any other changes in the circuit characteristics of Fig. 113. The current for an impressed electromotive force of 10 volts is shown for each of these assumed resistances for a range of frequencies from 100 to 600 cycles per second. From this may be gathered that other things being equal, the higher the resistance the flatter will be the curve. Expressed in another way, the critical frequency becomes less sharply defined as resistance is increased. In radio parlance, the lower the ohmic resistance, other things being equal, the more sharply tuned the circuit. At the resonant frequency the ohmic resistance constitutes the only opposition to the current flow and, as will be seen from the several curves, the currents at resonance are inversely proportional to the resistance. For all other frequencies, however, the reactance of inductance and capacitance is felt in increasing degree as resonance is departed from, and for frequencies widely removed from resonance the reactance, rather than the resistance, becomes the main controlling factor.

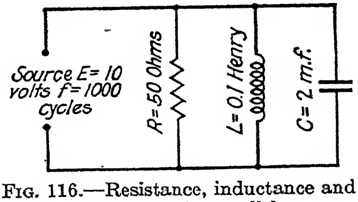


FIG. 116.—Resistance, inductance and capacitance in parallel.

Up to this point only circuits containing resistance, inductance and capacitance in series have been considered. Similar treatment may be accorded to multiple arrangements, such as that shown in Fig. 116 where the same circuit characteristics as those of Fig. 113 are connected in multiple instead of in series. As before, the impressed electromotive force will be taken at 10 volts. The frequency in this case will be considered as 1,000 cycles.

Here the voltage across the terminals of each branch is the same. The current in each branch may be determined separately

and the three combined vectorially to determine the amount and phase of the total current. Evidently, then,

$$I_R = \frac{E}{R} = \frac{10}{50} = 0.2 \text{ ampere.}$$

$$I_L = \frac{E}{2\pi fL} = \frac{10}{6.28 \times 1{,}000 \times 0.1} = 0.016 \text{ ampere.}$$

$$I_C = \frac{E}{1/2\pi fC} = 2\pi fCE = 6.28 \times 1{,}000 \times 2 \times 10^{-6} \times 10 = 0.126 \text{ ampere.}$$

I_R will, of course, be in phase with the impressed electromotive force E; I_L will lag by 90 degrees and I_C will lead by the same amount. The vector diagram appears then as in Fig. 117.

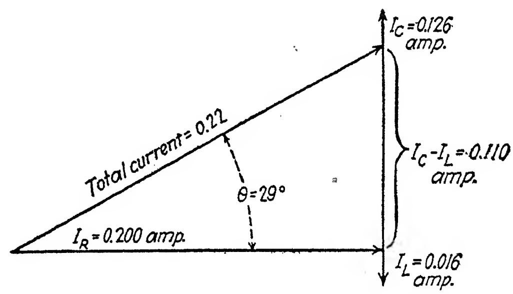


FIG. 117.—Vector diagram for current flow in circuit of Fig. 116.

From this it is evident that the total current of the three paths combined is

$$I_{\text{total}} = \sqrt{I_R^2 + (I_C - I_L)^2} = \sqrt{0.20^2 + 0.11^2} = 0.228 \text{ ampere.}$$

This current is leading the impressed electromotive force by an angle θ whose tangent is 0.110/0.200, or 0.55. The angle of lead θ is, therefore, nearly 29 degrees.

In the foregoing examples pure resistances, inductances and capacitances have been assumed. This is a condition not possible to attain. For instance, there can be no inductance without some resistance, and there is always some loss in the iron cores of coils due to magnetic hysteresis. Again, there is no such thing as a perfect condenser. Always there is some slight loss of energy within the dielectric itself (dielectric absorption) so that all of the energy used in building up the condenser charge is not restored on the discharge. All of these losses plus the actual work done outside the circuit, as in moving the

air surrounding a telephone diaphragm, for instance, do not contribute to the reactance, but, in one way or another, they generate heat and, therefore, act in phase with the true ohmic resistance. All of them together make up the apparent resistance indicated on the diagrams of Figs. 97 and 105.

Importance of Inductance and Capacitance.—In the earliest days of the telephone the action of capacitance and inductance on the rapidly varying voice currents was not appreciated at all. It was not understood, for instance, why the conductors in the cables then available or the windings of electromagnets offered any other impedance to the flow of voice currents than that of their mere ohmic resistance. Later, as the reasons began to be understood, inductance and capacitance were looked upon merely as necessary evils to be avoided wherever possible. Now, however, due to ever increasing knowledge of electrical and acoustical phenomena, these heretofore objectionable characteristics are being more and more turned to useful account. By an intelligent application of their properties, results are being accomplished which, without them, would be quite impossible.

To illustrate: The fact that the reactances of inductance and capacitance may be made to annul each other is one of far-reaching importance in telephone transmission. As already stated, every telephone line is inherently possessed of some electrostatic capacity, the two wires of the line acting as the plates and the intervening insulation as the dielectric of a condenser. The effect is that of a condenser, or, more properly, an infinite number of infinitely small condensers, connected across the line. Some of this "distributed capacity" is unavoidable and, in so far as it exists, it is detrimental to telephone transmission, tending to shunt the voice currents and prevent them from reaching their destination. The harmful effect is not only one of attenuation but of distortion as well.

As early as 1893, Oliver Heaviside, a distinguished British mathematician and physicist, suggested the possibility of overcoming the bad effects of this inevitable line capacitance by the introduction of the right amount of inductance. The proposition was somewhat akin to overcoming the effect of one poison by the application of another, either alone being poison, but the two combined being harmless. Nothing came of Heaviside's proposal for a number of years, probably because

there were few with sufficient mathematical knowledge to really understand his meaning.

In 1900, however, Professor Michael Idvorsky Pupin of Columbia University developed the mathematical theory, scarcely more than hinted at by Heaviside, and showed how the effects of distributed capacity in a telephone line could be largely mitigated by the introduction at stated intervals in the line of coils of predetermined inductance. Figure 118 shows the general scheme of a loaded telephone line, it being understood that the condensers are not actually present but are shown merely to simulate the capacity inherent in the line itself. This whole matter of the "loading" of telephone lines will be dealt with more extensively in another chapter and is mentioned here only as a striking example of one of the many important ways in which a

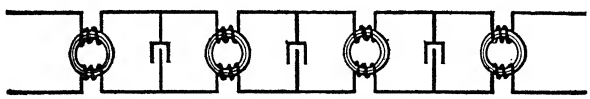


Fig. 118.—Principle of loaded line.

proper understanding of the combined effects of resistance, inductance and capacitance may be made to serve the purposes of telephony in the transmission of the very complex multi-frequency currents from one point to another.

As another example, the general subject of electrical wave filters, due principally to the work of George A. Campbell, may be briefly mentioned. Broadly speaking, an electrical wave filter is a device for differentiating between alternating current waves of different frequencies in much the same manner as wire screens of different sized mesh may be used to sift out and separate particles of desired sizes from a heterogeneous mass of sand and gravel.

As a very simple case, an impedance coil of low resistance and high inductance may be considered to act as a filter. Direct and low-frequency currents will pass through it with comparative freedom, while high-frequency currents will not. Here, however, there is no sharply defined dividing line between the currents passed and those barred. A tuned circuit, such as that of Fig. 113, may be made to show a much sharper line of demarcation. In this case the circuit, considered as a filter, will pass freely a narrow band of frequencies at or near the point

of resonance and offer great opposition to the passage of all others.

Campbell, however, carried the design of the electrical wave filters much further, so much further, in fact, that neither of the simple examples just cited would conform to the now generally accepted meaning of the word "filter." By the combination of carefully predetermined condensers and inductance coils in definite kinds of networks, he produced wave filters which will pass freely, and with about equal efficiency, all currents of a desired band of frequencies and effectively bar the passage of all others.

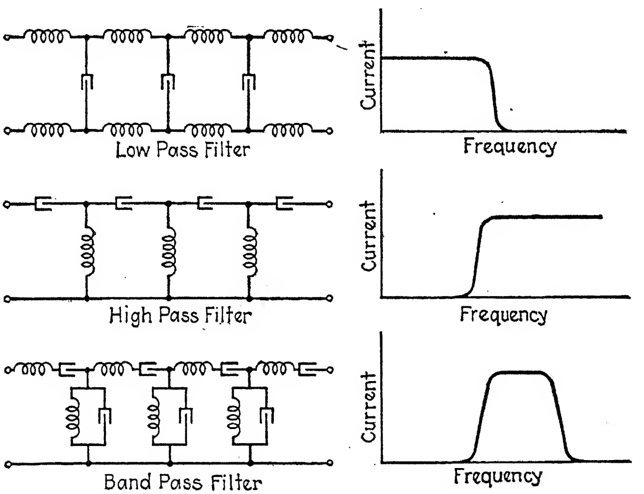


FIG. 119.—Filter circuits and their characteristics.

As examples, a typical filter of what is known as the "low-pass" type is shown with its characteristic performance diagram in the upper part of Fig. 119. This, as its name indicates, will pass, with little attenuation, all frequencies below the desired "cut-off" frequency and affect practically complete attenuation of all frequencies above. A somewhat different arrangement gives the opposite result. The center diagrams of Fig. 119 are typical of the so-called "high-pass" filters, which pass all frequencies above the cut-off and bar all those below it. The lower diagrams of this figure are typical of the "band-pass" filters. These transmit frequencies within a certain range but restrict the passage of frequencies above and below that range.

Typical characteristics of a band filter and of a resonant circuit are given, for comparison, in Fig. 120.

Networks of resistance, self-induction and capacity elements are not confined to filters. Often it is desired closely to simulate the characteristics of an actual line, and a proper network results in what is called an "artificial line." Again, networks are used in a corrective way to restore wave forms, which have been

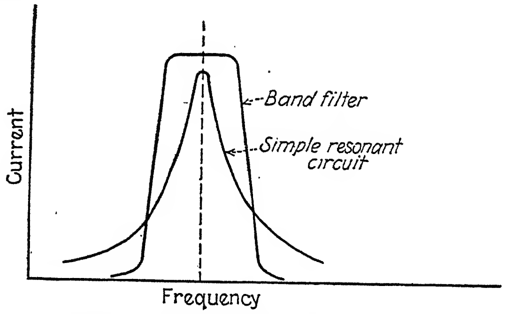


FIG. 120.—Characteristics of filter and simple resonant circuit compared.

somewhat distorted in transmission, to approximately their original form, or willfully to distort them if a good purpose can be served by so doing. Thus, if a given line has a tendency to attenuate the higher frequencies more than the lower, it is possible to provide a network at the incoming end which will act in exactly the opposite way, with the result that the distortion of the line is corrected in the network.

The complete consideration of various networks, whether for filters or for other purposes, often involves mathematical analyses of a nature that far transcends the scope of this work.

CHAPTER IX

VACUUM TUBE THEORY[1]

The term "vacuum tube" as used in this chapter refers to those highly exhausted bulbs which depend for their action on the phenomenon known as the "thermionic emission" from heated electrodes. This distinguishes them from other kinds of vacuum tubes, such as the well-known Geissler and Crookes' tubes, which depend upon the ionization of residual gas, and also from ordinary incandescent lamps which employ heated filaments as sources of light. The thermionic vacuum tube did not come into existence as a useful instrumentality in electrical communication until the present century was well under way. Notwithstanding its very recent arrival, it has already proved of vast usefulness in telephonic and telegraphic wire communication and has made possible the new art of radio broadcasting. More than this, it has proved a most useful tool in the hands of the physicist and, as such, has gone far in helping mankind to a better understanding of the fundamental nature of things about him. It is perhaps the most revolutionary physical thing of modern times.

As early as 1884, Edison, observing peculiar effects in the discoloration of the inner walls of certain incandescent lamps, was led to investigate the cause. In doing this he included a metal plate within the exhausted space of an ordinary carbon filament incandescent lamp. This was insulated from the filament and connected to a third terminal outside the bulb. He observed that when the added plate was connected, as in Fig. 121, by an external conductor with the positive terminal of the filament while it was incandescent, a current would flow through this conductor. This evidently indicated a flow of

[1] For more complete discussions of vacuum tubes than the scope of this work permits reference is made to "The Thermionic Vacuum Tube," by van der Bijl; "Principles of Radio Communication," by J. H. Morecroft; and "Thermionic Vacuum Tubes and their Applications," by R. W. King. I have drawn freely on these and many other sources in the preparation of this chapter.

current across the space within the vacuum between the plate and the filament. When, however, the added plate was connected to the negative terminal of the filament, no such current was observed. This was known as the "Edison effect." It was for a long time allowed to pass as an interesting phenomenon for which there was no adequate scientific explanation. The fact remains, however, that Edison produced at that early date what we now know as the "two-electrode thermionic vacuum tube." It was ahead of its time, for there was not then sufficient knowledge to account for its action nor to appreciate its possibilities. Moreover, electrical communication had not then developed far enough to need it.

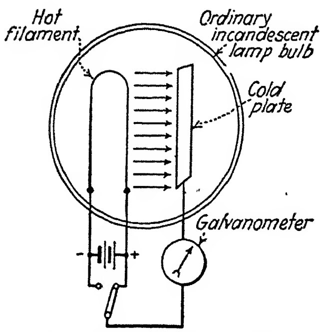


FIG. 121.—The Edison effect.

To understand the action of the thermionic vacuum tube, we must go further than the physical conception of matter as it was taught a generation ago and inquire concerning the more modern view. The atom, formerly considered the smallest possible subdivision of matter, is now believed to be made up of much smaller units called *protons* and *electrons*. These move about among themselves in the relatively large space of the atom in some such manner as the bodies of our solar system move among themselves in stellar space. We have then, in the atom, a massive nuclear sun (the proton) around which planets (electrons) move.

It is customary to think of these protons and electrons as tiny corpuscles or particles, each constituting an electric charge. The proton charges are positive, the electron charges negative. Very recent discoveries, however, seem to indicate and even prove that the electrons and protons are in the nature of waves instead of particles; just as X-rays are waves and not particles. And so at this moment science is confused, some experiments pointing to the corpuscular theory; others, equally weighty, pointing to the wave theory. Strange as it may seem, each of these views is supported by such important evidence that, although they are apparently inconsistent, both are being accepted. Physicists today are seeking further knowledge,

and, so fast is science moving, the latest discoveries tend toward reconciling the apparent inconsistency rather than overthrowing either of the theories.

Whatever the nature of the electron, it is known that it is, or that it carries, the smallest charge of electricity so far isolated. *A stream of electrons, therefore, constitutes a current of electricity.*

We may now go back to the question of thermionic emission. It has been known for nearly two centuries that a hot body charged with electricity will lose its electric charge more rapidly than a cold one. No theory which afforded a complete explanation of this was put forth until the modern conception of the constitution of matter, just alluded to, was developed. Under this conception it is assumed that in a conductor some of the electrons are free to move about, and that the average velocity with which they are moving increases when the absolute temperature of the conductor increases.[1]

At the surface of the conductor there is surface tension—an electric field—which ordinarily prevents these moving electrons from escaping into outer space. The electrons at any given temperature have individual velocities which differ greatly among themselves. As the temperature of the conductor is increased, these velocities increase and finally some of the swifter moving electrons acquire enough velocity to break through the surface field and to escape. This, very broadly, is the conception of the thermionic emission of electrons from a conductor.

When we take a negative charge from a neutral conductor, we leave that conductor positively charged. Therefore, when an electron, which, it is to be remembered, is a negative charge, is thrown off from a neutral conductor, it leaves the conductor positively charged. If there is no other conductor in the immediate vicinity, the electron which we have just conceived as

[1] Since thermionic emission is directly related to the amount of heat in the conductor from which the emission takes place, it is more directly related to a scale of temperature which starts at a condition of the theoretical absence of all heat rather than at any of the arbitrary starting points, such as the melting point of ice. It is known from theoretical considerations that a body from which all heat has been extracted would have a temperature of 273 degrees below zero Centigrade. This, therefore, is "absolute zero." The scale of absolute temperature above this point is called, from its originator, the "Kelvin scale," and is expressed in "degrees Kelvin" or merely "degrees K." As an illustration, the temperature of melting ice is 273° K., that of boiling water 373° K.

breaking away from the conductor will proceed on its journey for a short distance, and then, its kinetic energy becoming exhausted by collisions or in other ways, it will be drawn back to the conductor, due to the attraction between dissimilar charges. This process undoubtedly goes on all the time in an ordinary incandescent lamp while its filament is heated.

If, however, we assume that there is another conducting body near by, with a positive charge, the free electron may have sufficient velocity to project itself far enough into the field of attraction of this other conductor to continue on its way and find lodgment in this new body. If we have a number of electrons breaking loose from the heated body and passing to the near-by positively charged body, we may call it a stream of electrons, or, in other words, an electric current passing between the two. If, on the other hand, the near-by body has a negative charge, it will repel instead of attract the fugitive electron and thus aid in its return to the heated conductor whence it started. In this case there will be no stream of electrons across the space and, consequently, no current.

In the light of this theory the explanation of the Edison effect is now apparent. When the added plate within the tube (Fig. 121) was connected to the positive terminal of the incandescent filament, it was, of course, positive with respect to the filament as a whole, on account of the drop of potential along the filament from its positive to its negative terminal. There was, therefore, a flow of electrons from the filament to the plate. When the plate was connected to the negative side of the filament, it was, by the same reasoning, negative to the filament as a whole, under which condition the electrons emitted were driven back to the filament, and, consequently, there was no flow of current.

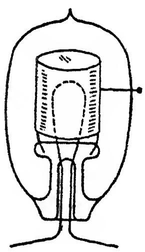


Fig. 122.—Fleming valve.

It was not until 1904 that Professor J. A. Fleming of London made use of this effect. One form of his tube, indicated in Fig. 122, consisted of an ordinary carbon filament incandescent lamp within the glass bulb of which he included a metal plate bent into the form of a hollow cylinder surrounding but not touching the filament. This, it will be seen, contains the same essential elements as the modified lamp which Edison was using when he discovered the Edison effect. Fleming turned his

device to useful account while, for the reasons stated, Edison did not.

Fleming called his tube an "oscillation valve," and it has since been widely known as the "Fleming valve." Such tubes, having two electrodes only—an electron-emitting filament and an electron-receiving plate—are now commonly known as "two-electrode" tubes. The name "valve," however, is apt. Such tubes conduct current in one direction and not in the other, in much the same manner that a one-way valve in a water pipe will permit the flow of water in one direction only.

The action of the two-electrode tube, with respect to its unilateral or unidirectional conductivity, may be made clear from a consideration of the two diagrams of Fig. 123, if it is not

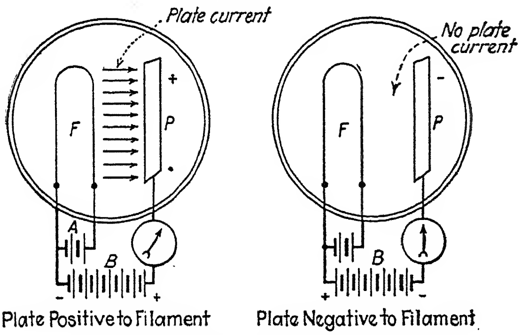


FIG. 123.—Action of two-electrode tube.

clear already from what has been said of thermionic emission. Briefly, when the filament is heated by a flow of current through it, as from the battery A, it throws off electrons into the surrounding space. What happens to these electrons depends largely upon the relative potentials of the filament and the plate.

If, as in the left-hand diagram, the plate is held positive with respect to the filament, as by battery B, it will attract some or all the vagrant electrons to it, and the resultant stream of electrons will constitute a current of electricity across the gap, which will be indicated by the galvanometer needle connected in the external circuit. If, on the other hand, the battery B is reversed so that the plate is made negative to the filament, the electrons will be repelled by it and, therefore, assisted back to the filament. In this case there is no electron stream across the space and no flow of current.

Here it is well to call attention to an apparent inconsistency. We say that an electron stream is a current and we show that in a vacuum tube the electron stream passes always from the negative to the positive electrode. This is exactly contrary to the conventional idea as to the direction of current flow. This is unimportant, since the conventional idea was based on a mere assumption. It is likely, however, to cause confusion. If the old conventional positive-to-negative idea of direction is to persist, as it probably will, then, in order to be consistent, we must say that the current flow is in the opposite direction from the direction of the electron stream. In order to avoid confusion, as far as possible, the arrows in the diagrams of this chapter will indicate the direction of the electron stream, ignoring the conventional practice of indicating current flow from positive to negative.

It was the addition of another electrode, the "grid," that gave the vacuum tube its final impetus and made possible the recent astounding development of both wire and wireless communication. This was done in 1906 by Dr. Lee De Forest, who, as a result, must be given credit for one of the greatest inventions of modern times. De Forest called his tube the "audion." In its various forms it has been given many other names. A generic name applicable to all forms is the "three-electrode vacuum tube."

The third electrode in the form of an open mesh or grid is placed within the tube, usually in such position as to lie between the filament and the plate and thus in the path of the stream of electrons emanating from the filament and flowing to the plate. It usually lies much closer to the filament than to the plate and from its general form takes the name "grid." It is insulated from both filament and plate and connected to a separate terminal outside the tube. The tube, therefore, has four terminals: two for the filament, and one each for the plate and grid.

The general appearance of such a three-electrode tube is shown in Fig. 124, this being of a form of tube manufactured by the Western Electric Company. It is largely used for telephone repeaters as well as for many other purposes of wire communication. There are, of course, numerous variations from this type, in size, form and detail, due to the varying purposes for which they are to be employed and the varying methods used in their manufacture. They all consist essentially of the three elements,

filament, plate and grid enclosed in a highly exhausted bulb, and they may all be represented, for the purpose of this chapter, by such a conventionalized symbol as that of Fig. 125.

If we think of the two-electrode tube as a valve, then we may consider the grid as a marvelously delicate control for this valve. The most minute amounts of energy applied to the grid serve to control the very much larger energy of the plate circuit. A preliminary idea of how the grid exerts this control may be had by considering Fig. 126. In this the filament is supposed to be heated by a battery not shown. The plate is held at a positive potential with respect to the filament by the battery B, and the grid

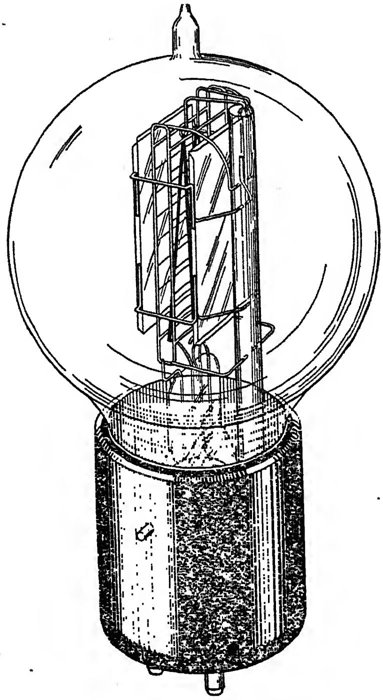


Fig. 124.—Three-electrode vacuum tube.

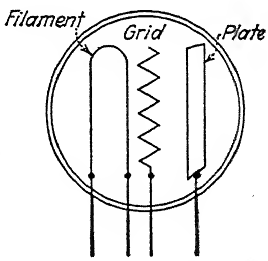


Fig. 125.—Conventional representation of three-electrode tube.

may be made either positive or negative to the filament to any desired degree by connecting it to different points along the battery C. If the potential on the grid is such that the electric field is not modified, the electron stream would merely pass through the open spaces of the grid with little change. With the grid held negative, however, it is obvious that it tends to drive the electrons back to the filament, thus decreasing the plate current. If sufficiently negative it, will establish a field through which the electrons cannot pass,

thus entirely stopping the plate current. On the other hand, a positive grid will assist the plate in attracting electrons, thus increasing the plate current. Clearly, according to its potential, the grid will exert a marked influence on the plate current. The utility of the three-electrode tube depends on the extreme sensitiveness of plate current to changes in grid potential.

To obtain a somewhat more complete understanding of this control, we may consider the conditions in the space between the filament and the plate, without regard, at first, to the action of the intervening grid. When the filament is heated above the temperature of emission, there is an atmosphere of electrons

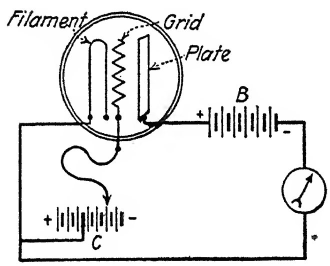


FIG. 126.—Action of grid in three-electrode tube.

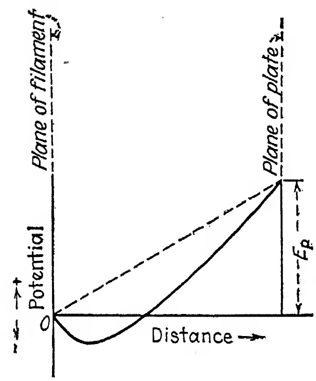


FIG. 127.—Potential across space as affected by space charge.

in its vicinity; and when the plate is held at a relatively positive potential high enough to draw some of these electrons to it, this atmosphere extends to the plate. This atmosphere of electrons constitutes a negative charge in the space between the electrodes, which is called the "space charge." The effect of this space charge on the potential at different points between the electrodes may be considered in connection with Fig. 127, where it is assumed that the filament and plate are in the form of planes facing each other and of dimensions large compared to the distance between them, this arrangement giving a uniform electric field. It is also assumed that the filament is at a zero potential and that the plate is maintained at a positive potential E_p above it. If there were no electrons within the space, the potential across the space from filament to plate would rise in a uniform gradient, as indicated by the dashed line. Obviously,

the presence of the electrons constituting the space charge will lower this potential, and the line representing its value at points across the space will assume some such form as that of the curved line in Fig. 127. This lowering of the line representing the potential will be more pronounced near the filament, since there the electron density is greatest.

Evidently, under the conditions of Fig. 127, an electron to get to the plate must pass through a negatively charged volume, and, of course, this negative volume charge repels it. Hence, to get through it must have enough kinetic energy due to its velocity to overcome this initial repulsion. This is the critical part of its journey, and any help or hindrance encountered here is likely to determine whether it proceeds to the plate or returns to the filament. It is here that the grid is placed to exert varying degrees of help or hindrance, according to its own potential. Expressed in another way, the function of the grid is to alter the potential gradient between the filament and plate, and thus exert an influence on the number of electrons that will pass through the intervening space.

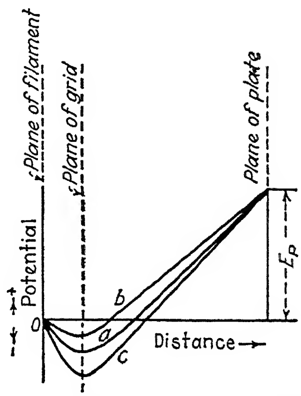


Fig. 128.—Potential across space as affected by potential of grid.

Figure 128 shows a similar diagram with the grid added. Of the three potential curves shown, the one marked *a* corresponds to the lower one in Fig. 127 and represents the potential condition across the space with no grid present, or, more properly, with no potential on the grid which would modify the gradient. The line *b* indicates the potentials which would exist were the grid raised to a potential higher than that due to the normal space charge in the plane of the grid; and, similarly, the line *c* represents the gradient as it would exist if the grid were given a potential lower than that due to the normal space charge in its plane.

It is obvious that the flow of electrons from filament to plate will be very sensitive to such modifications in this potential gradient as may be produced by charges on the grid. When the grid potential is raised, many of the electrons which would

otherwise have fallen back to the filament will be attracted to its vicinity. Having arrived in this vicinity, some of them may be attracted to the grid and thus form a small "grid current," but most of them, on account of the higher potential of the plate, will continue on to it. When, however, the potential of the grid is lowered, it will repel more of the electrons back to the filament, and only those which have comparatively large initial kinetic energy will be able to get through the low-potential field of the grid and pass on into the field of attraction of the plate.

The vacuum tube, owing to the characteristics which have so far been only broadly dealt with, finds a remarkably wide range of usefulness in the communication arts. It is used as a rectifier of alternating currents; as a detector of extremely small alternating voltage variation; as an amplifier of currents or voltages; as a modulator of alternating currents; and as a generator of current oscillations. Since the advent of the three-electrode tube, the two-electrode type is now used only as a current rectifier. The only field in which the usefulness of the two types of tube may possibly overlap is in detection, and in this the three-electrode tube is more efficient. The two-electrode tube cannot generate oscillations, amplify or modulate.

Before discussing any of these practical applications, some consideration will be given to such of their underlying characteristics as are common to both two-electrode and three-electrode tubes. These characteristics are both structural and operational. Among the important structural features are the degree of vacuum to which the tubes are exhausted, the material and dimensions of the electrodes and their relative spacing within the tube. These, of course, all combine to determine the operational characteristics, such as the effect of variations in electrode temperature or in electrode potential on the current passing through the interelectrode space, the power capacity, the efficiency, the constancy and, finally, the life of the tubes.

Vacua.—A fundamental reason for the exhaustion of vacuum tubes is to avoid the oxidation or other chemical change which would promptly burn up the filament were it heated in air. But aside from this there are two principal ways in which the presence of any considerable amounts of gas in vacuum tubes would produce undesirable effects in their operation. These van der Bijl considers under two headings, "volume effects," referring to the volume of the evacuated space, and "surface

effects," referring to the surfaces of the elements and of the inner walls of the vessel.

Because of the almost complete lack of carriers of electricity in the form of free charges, gases under ordinary conditions are practically non-conducting. The carriers of electricity in gases consist of electrons and positive and negative "ions." An ion is essentially a charged atom and may be either negative or positive. A negative ion results when an atom for some reason acquires one or more extra electrons; a positive ion when an atom loses one or more electrons.

One way to produce ions in abundance is to bombard a gas with high-speed electrons or ions. These, on collision with the molecules, break them up into positive and negative ions, and the process is known as "ionization by collision." There are always present in gases a few ions resulting principally from radiations of one kind or another. If sufficient electric stress is applied to the gas, these existing ions acquire the velocity necessary to ionize by collision. Each collision produces two new ions, causing the total number to rise at a perfectly enormous rate and making the normally insulating gas a good conductor. This is the ordinary phenomenon of the electric spark. If the gas pressure is reduced, which is equivalent to removing some of the molecules, the free ions travel farther between collisions. Hence a smaller electric stress is sufficient to give them the velocity necessary to ionize by collision. In other words, at low pressure less voltage is required to give the same length of spark, or the same voltage will produce a much longer spark, as in the well-known Geissler tube. Further, since there are fewer molecules involved and, therefore, fewer ions, the conductivity will not rise so high and the whole phenomenon will be less violent. The nature of the electrodes enters into both these phenomena to a certain extent as well as the nature of the gas.

If the pressure is still further reduced, a condition will be reached where ions will travel the whole distance between the electrodes without making a collision. Also, the number of naturally produced ions will be negligible, and conduction will cease entirely as far as the gas is concerned. The only way to produce conduction at this stage is to cause the electrodes to emit ions or electrons, and when one electrode is caused to emit electrons by heating, we have the "vacuum tube."

It is important in vacuum-tube operation that the current flow be due solely to thermionic emission, and in no appreciable degree to conduction through gas. If the evacuation of the tube is not sufficiently perfect, some of the molecules of the residual gas will be struck by the electrons emitted by the filament and the gas will be ionized, thus becoming a conductor on its own account. As a result of this the tube may acquire characteristics quite different from those it would have if the current through it were due to pure thermionic emission.

The requirement from this standpoint is, therefore, that the tube be evacuated to such an extent that there will be practically no collisions between the emitted electrons and the molecules of the residual gas. The residual gas then plays no appreciable part in the current flow between the filament and the plate.

Even with the highest vacuum attainable there are, in the remaining gas, something like one hundred million molecules of gas per cubic centimeter; and in ordinary tubes, such as those used in commercial radio receiving sets, they are about one thousand times as numerous. This, it must be remembered, is after the tube has been exhausted to a pressure of perhaps one billionth of an atmosphere.

Taken by themselves, these figures might seem to indicate that there is an appalling crowding together of molecules within the space we call a vacuum. Actually, the molecules are so small that, even with this number of them present, the chance is very remote that any one of them will be struck by the horde of very much smaller electrons passing through the same space. In a properly evacuated tube the chance is small enough to be negligible.

From the standpoint of the volume effect, the main point is, as far as possible, to get rid of all the gas within the tube so as to avoid ionization by collision. If there is any considerable amount of residual gas, the splitting up of its molecules, under the impact of the emitted electrons, may so alter the distribution of the electric charges within the electrode space, and the flow of current through it, as to partially or wholly obscure the desired action of the tube.

The effects of residual gas on the surface of the filament are of no less importance than the volume effects. It is found, for instance, that the presence of a very small amount of water vapor will so alter the surface of a tungsten filament as to enor-

mously decrease the electron emission. Certain oxides act in the same way and are said to "poison" the filament surface.

Without attempting to discuss all of the bad effects of residual gas, it may be said in general that the higher the vacuum the more consistently does the actual experimental and practical performance of the tube agree with the predictions from purely theoretical considerations of thermionic emission.

In practice it is not sufficient merely to extract, as far as possible, all the gas from the volume of the tube. Care must also be taken to free the electrodes and the walls of the bulb from the gases which are occluded on or beneath their surfaces. Unless this is done, these occluded gases would subsequently be driven off, thus spoiling the vacuum and otherwise interfering with the operation of the tube.

In the manufacture of its vacuum tubes for the Associated Bell Companies, the Western Electric Company finds it necessary in order to secure both high and permanent vacua to employ the following process, of which only the barest outline may be given here:

As the last step before inclosing the electrodes and other metal parts within the glass bulbs they are given preliminary heat treatments in vacuum furnaces, and also in a hydrogen atmosphere, in order to cleanse the metals internally and drive off all volatile matter from their surface. The temperature to which these parts are subjected in this treatment is about 1000° C.

After placing the pretreated parts within the bulbs, the air is exhausted by the combined action of oil pumps, mercury pumps and liquid air traps. The purpose of the liquid air trap is to prevent the mercury vapor of the pumps from finding its way into the tubes by freezing it fast to the walls of the trap. During the process of pumping the tubes are inclosed in an electric oven in which they are maintained at as high a temperature as possible without softening the glass. Sufficient current is then passed through the filaments to heat them and cause free electronic emission, while at the same time the plate potentials are raised to high values. As a result the plates receive an intense electron bombardment, which raises them to a temperature of over 800° C. During this process the pumping is continued so that any occluded gases or other matter within the tubes which had not been driven off by the preceding heat treatments may be withdrawn. As a final operation a very small amount of one of the alkali metals is

vaporized within the tube, producing the mirror-like coating often observable on the inner wall. The purpose of the introduction of this vapor is twofold: To react chemically with any water vapor or other oxygen content which might otherwise remain free in the tube, and to "bury" under the mirror coating any other gases which might remain on the inner surfaces.

The tube is then "sealed off" from the pump. The pressure within it, as a result of these operations, is in the neighborhood of one billionth of an atmosphere.

Cathode Material.—There are a number of ways in which an electrode *in vacuo* might be heated in order to secure an emission of electrons from it. It might, for instance, be subjected to bombardment by a stream of electrons from another source. In fact this sometimes occurs in vacuum tube operation as a secondary and usually undesirable effect. The most convenient way, however, and the one almost universally employed, is to make the cathode in the form of a filament and to heat it by passing an electric current through it. The current supplied to the filament for this purpose is called the "power current," and it is one of the items of expense in the operation of vacuum tubes. The emission of electrons is vastly greater for high temperatures than for low, increasing at a very much higher rate than the power required to heat the filament. Considered from this standpoint of power consumption only, economy would dictate the running of the tubes at as high temperatures as possible.

There are, however, other considerations of economy which tend in the opposite direction. The life of the filament, and therefore of the tube, is very much shortened by running at high temperatures. The replacement of tubes becomes an important offsetting element of cost if their lives are unduly shortened. Evidently, considering these two factors only, the economic temperature at which to run the filaments will be the one at which the sum of the power costs and the tube replacement costs is a minimum. The choice of filament material has an important bearing on both the current consumption and the life of the tubes and also, as will be shown, on the general performance characteristics of the tubes.

There are, of course, other questions than those of economy to be considered. It is of prime importance that the tube be constant in its characteristics—that it operate for long periods of time without noticeable change in its behavior. This again

points to the desirability of operating at relatively low temperatures, since it is found that greater constancy can be secured by so doing.

The emission of electrons from heated conductors not only varies widely for different kinds of cathode material but also is extremely sensitive to the condition of the cathode surface. These and the other considerations just mentioned have naturally led to extensive experiments to determine the best possible material for the filament and the best kind of treatment for its surface.

For an escaping electron to break through the surface tension of the filament conductor it is necessary for it to do a certain amount of work. This amount of work is called the "electron affinity" or the "work function" and it varies with the substance of which the filament is made and with its surface condition. The lower the work function the smaller will be the velocity which the electron must attain in order to break through, and, therefore, the lower the temperature at which the required emission can take place. Materials or surfaces having low work functions thus permit emission at relatively low temperatures with the double advantage of low cost for power and of long life.

The choice of filament materials lies among three general types:

1. Pure metal filaments made of such refractory materials as tungsten, tantalum, molybdenum and osmium. These have relatively high work functions, but to offset this they are so highly refractory as to enable them to be worked at sufficiently high temperatures to obtain copious emission.

2. The so-called "thoriated tungsten filament" accredited to Langmuir. These are of tungsten in which a very slight amount of thorium is dissolved. Here, as will be shown, the electron emitting surface becomes one of thorium rather than of tungsten. It has a much lower work function and emission occurs at temperatures far below those at which it occurs from a pure tungsten surface.

3. The oxide coated filament accredited to Wehnelt. In these the oxides of certain alkaline earth metals are applied as a coating to a wire or ribbon of some refractory metal, such as platinum. These cannot be operated at such high temperatures as the pure metal filaments, but, having a much lower work function, ample emission occurs at lower temperatures.

Of the pure metal filaments tungsten has been the most widely used. As the light giving element of incandescent lamps the tungsten filament has demonstrated its ability to stand high temperatures for long periods of time. But for most thermionic vacuum tube uses it is neither as efficient nor as durable as either of the other types, and for these reasons has largely been superseded by them. It is not as efficient as the others because, having so much higher electron affinity, more energy is required in heating current to develop a given amount of emission. It is less durable, particularly in tubes which require thin filaments, because of the higher temperature at which it must operate. The rate at which a filament volatilizes increases with the temperature. As the pure metal filament volatilizes, it gets thinner and this acts detrimentally in two ways. It increases the filament resistance thus altering its operating characteristics, and it causes the filament to rupture easily, thus destroying the usefulness of the tube. The operating temperature range for pure tungsten filaments is in the neighborhood of from 2400 to 2600° K., depending on the diameter of the filament and on the thermionic efficiency desired.

The life of the filament depends on the temperature and upon the initial thickness. The thicker the filament the longer the life for a given temperature. These considerations lead to a favorable position for the pure tungsten filament where large amounts of power are involved. The large power consumption permits a thick filament; a thick filament permits operating at high temperatures with a reasonable life; and the higher the temperature the greater the thermionic efficiency. For these reasons the pure tungsten filament has survived and is largely used in power tubes.

The thoriated tungsten filament is made of tungsten from which has been removed, as nearly as possible, all impurities except about one-half of one per cent of thorium oxide and a small amount of carbon. As this filament is heated, minute quantities of metallic thorium come to its surface, resulting in an exceedingly thin layer of that metal on the surface of the tungsten. Such a filament gives copious emission of electrons at a temperature of about 1500° K., far below the emission point of tungsten. This, together with other facts, seems to indicate that the emission is wholly from the thorium coating, the tungsten acting merely to develop the heat. As the electrons evaporate from

the surface of the thorium, other electrons from within the body of the filament come to the surface to replace them.

When the thoriated tungsten filament is raised a few hundred degrees above its normal operating temperature, its thorium coating disappears, and such additional thorium as comes to the surface also evaporates at once. The removal of the thorium coat in this way is called "deactivation" and the emission from the filament becomes that of pure tungsten, which means practically no emission at the normal operating temperature of the thoriated filament.

Again, a thoriated filament under certain conditions may have its coating removed or injured by the bombardment of positive ions when the residual gas in the tube becomes ionized, so that proper emission cannot occur.

When a thoriated-filament vacuum tube has been "deactivated" in either of these ways, it usually may be "reactivated" by a rather simple treatment. The filament is first raised for a few seconds to a very high temperature (about 2800° K.) which thoroughly cleanses its surface. It is then subjected to a temperature of about 2100° K. for several minutes. During this time the coating of thorium is replenished from the supply within the filament. In this operation no voltage is applied to the plate or grid. A thoriated filament may be reactivated several times in this manner before the supply of thorium within the filament is finally exhausted. This is one of the advantages strongly urged in favor of the thoriated type of filament.

The third class of filaments mentioned, the "oxide coated," is the one used almost exclusively for the purposes of wire telephony. As early as 1905 Wehnelt discovered that a speck of lime placed on a filament of platinum had a very high electronic emission at comparatively low temperatures. He later experimented with other alkaline substances and from his work the class of electrodes coated with oxides of the alkaline earth metals are known as the "Wehnelt cathodes."

The first oxide-coated filaments were extremely erratic, both as to life and performance. The coating material would flake off from the metallic body of the filament. It was not only impossible to secure uniformity of product in their manufacture, but such tubes as were made would not remain constant in their characteristics even during their comparatively short lives. The engineers of the Bell System recognized, however, that this

class of filaments possessed inherent advantages above all others, and by intensive research overcame these difficulties.

So successful has this development work been that tubes containing these filaments are now manufactured in large quantities with such uniformity and such constancy that they are completely interchangeable, even under conditions of most exacting requirements. Furthermore, durability has been so increased that instead of lasting less than 1,000 hours in service they now show average life of certainly more than 20,000 hours, and perhaps as much as 100,000 hours. Since a year contains only about 8,760 hours, there has not yet been sufficient time since their production to determine average life with accuracy.

Coupled with this improvement in durability there has been a large reduction in power consumption. As an illustration, the type of tube used in telephone repeaters has had its power consumption reduced from about nine watts to about two watts, the latter consumption being based on a much more conservative temperature rating than the former. As a result of these performances, the annual cost per tube socket for power and for tube renewals has been reduced from about $100 per annum to a small fraction of that amount.

The core of the oxide-coated filament as thus developed consists of a very thin ribbon of platinum-nickel alloy, which is twisted on itself so as to form a succession of curved surfaces well adapted to hold the oxide coating it is to carry. This filament ribbon serves to carry the current and generate the required heat but does not in itself furnish the surface from which the electronic emission occurs. The source of emission is the oxide coating, which consists of a mixture in paraffin of the oxides of barium and strontium. This is applied to the filament in successive very thin layers, each of which is baked on. Effective thermionic emission from filaments made in this way occurs at temperatures around 1100° K.

A general indication of the difference in temperatures at which the foregoing three types of filaments give effective emission is given by their appearances in normal operation: The pure tungsten filament works at a dazzling white heat; the thoriated tungsten at a yellow heat; and the oxide coated at a dull red glow, scarcely noticeable except in subdued light.

The amount of thermionic emission from a heated cathode depends on three factors: the area of the cathode surface, its tem-

perature and the electron affinity of the material of which it is made. When the conditions are such that all of the electrons emitted by the filament are drawn to the plate, the current so formed is called the "saturation current." No greater current can flow at that temperature because no more electrons are being emitted. The saturation current for a given filament temperature is, therefore, the total emission at that temperature. It is commonly measured in milliamperes. For purposes of comparing the

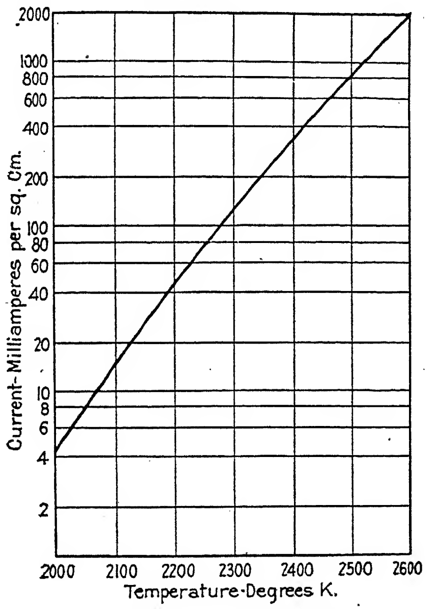


FIG. 129.—Emission from surface of pure tungsten filament at various temperatures.

relative thermionic efficiencies of different filament materials, it is often convenient to refer to their emissions in terms of saturation current per unit area of filament surface.

The emission per square centimeter of surface from a pure tungsten filament at different temperatures is shown in Fig. 129. This was derived from purely theoretical considerations by Morecroft using Richardson's equation. It is to be noted that the ordinates of this curve are plotted on the logarithmic scale in order to keep the diagram within reasonable length.

In the same manner, by using the power supplied to maintain the filament temperature, a curve may be drawn showing the

relationship between the maximum thermionic emission (saturation current) and the power required to maintain the filament at the desired temperature. This is an important relationship, and, fortunately, it is much easier to determine experimentally than the relation between temperature and emission, because the

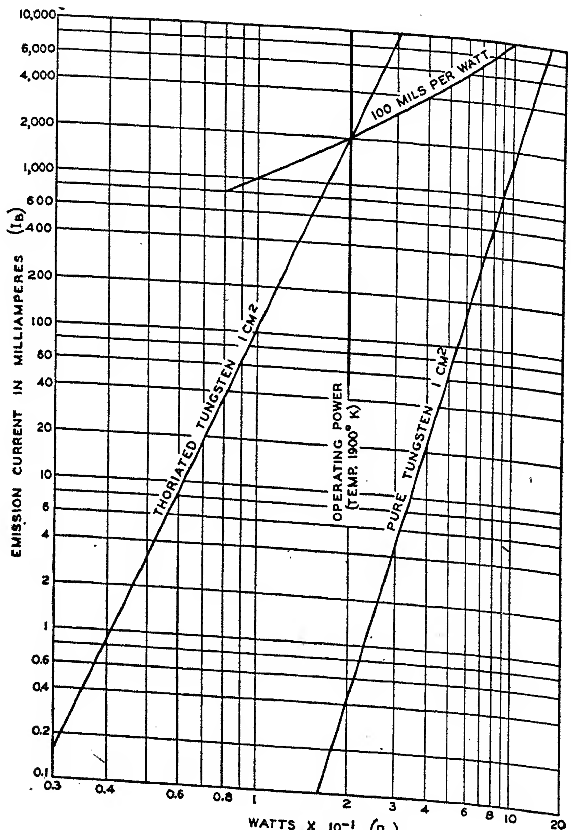


FIG. 130.—Power-emission chart for pure and thoriated tungsten filaments.

electrical power supplied to the filament is more readily measured than the temperature of the filament.

Such curves, as given by King, for the three types of filaments, pure tungsten, thoriated tungsten and platinum-nickel oxide coated, are shown in Figs. 130 and 131. The same curve for pure tungsten is given in each of these in order to facilitate comparison with the other types. In the coordinate paper on

which these charts were drawn the abscissæ are curved in order that the relation between the emission from the filament and the power supplied to it will be represented by straight lines. The curves in each case represent the emission from one square centimeter of filament surface.

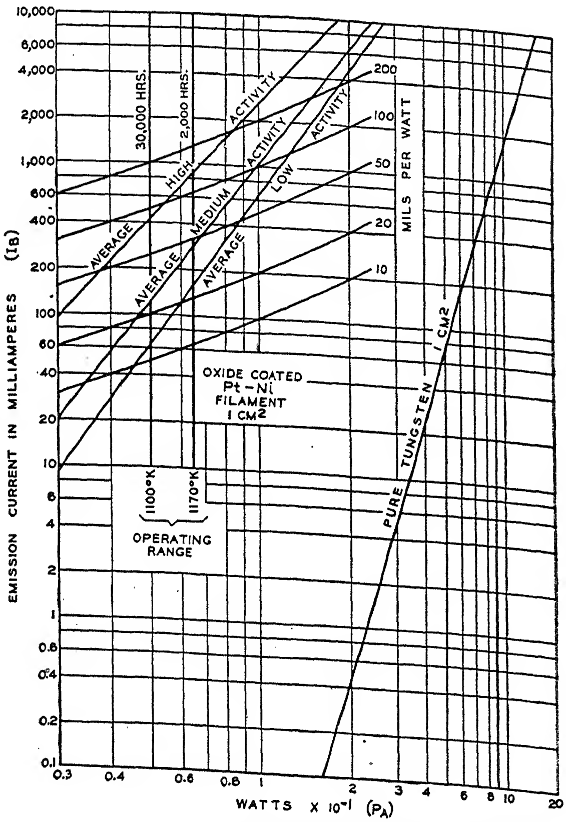


Fig. 131.—Power-emission chart for oxide-coated and pure tungsten filaments.

There is considerable variation in the emission of different samples of oxide-coated filaments. For this reason three curves are shown in Fig. 131 for this type, applying, respectively, to average filaments of low, medium and high activities. The efficiencies in milliamperes of emission per watt of power used to heat the filament are shown by the curved lines crossing each chart.

The heavy vertical line of Fig. 130 shows the power, with the corresponding temperature indicated that would have to be applied to a thoriated tungsten filament to secure an efficiency of 100 milliamperes per watt of heating current, the power being 20 watts and the emission 2,000 milliamperes. This would, however, produce a filament temperature of 1900° K., which, according to Morecroft, is about 400° above normal operating temperature. A similar chart by Morecroft shows that at 1500° K. the thoriated filament would consume about 16 watts and give an emission of nearly 500 milliamperes, an efficiency of about 30 milliamperes per watt. This is probably more nearly representative of safe operating conditions, such as would give an average life, counting on reactivation, in excess of 2,000 hours.

Compared with these figures the oxide-coated filament shows advantageously, particularly when a balance between the two conflicting requirements, long life and high efficiency, is considered. Thus, when compared on about the same (2,000 hours) average life basis, the oxide-coated filament requires about 6.5 watts per square centimeter, with a resulting temperature of only 1170° K. At this power consumption the low, medium and high activity filaments show efficiencies of about 25, 46 and 140 milliamperes per watt, respectively. When operated at the lower temperature of 1100° K., the oxide-coated filament shows an average life of 30,000 hours, with somewhat lower thermionic efficiencies; 12, 24 and 80 milliamperes per watt of heating current for the low, medium and high activity filaments, respectively.

As compared with these thermionic efficiencies of the thoriated and oxide-coated filament, it is seen that within the range shown on Figs. 130 and 131 the emission from a pure tungsten filament with a surface area of one square centimeter varies from practically nothing at a power consumption of 16 watts up to an emission of 1,000 milliamperes for a power consumption of 500 watts, the latter figures representing an efficiency of only 2 milliamperes per watt. It is to be remembered, however, that for power consumptions and higher temperatures beyond the upper range of these charts, the efficiency of the pure tungsten filament would considerably increase.

Tube Characteristics.—So far, in connection with Figs. 129, 130 and 131, we have been dealing with the total emission from the filament, this constituting the saturation current under the assumption that all the emission is being attracted to the plate, or plate

and grid. Vacuum tubes, however, usually operate at space-current densities well below saturation, and it is useful to consider their operating characteristics in that range. We may take the simplest case first, that of the two-electrode tube, or the three-electrode tube with the grid considered inactive.

If we keep the filament current constant so that variations of filament temperature will not enter into the action, we find that, between certain limits, the plate current will depend on the degree to which the plate is held positive with respect to the filament. These limits are: zero at the lower end, where the plate potential passes through zero and changes to negative, for, as we have seen, there can be no stream of electrons across the space when the plate is negative to the filament; and the saturation current at the upper end for here the plate current depends only on the electron emission. No matter how high the potential, within a reasonable limit, the plate cannot draw to itself more electrons than the filament is capable of emitting at a given temperature. Between zero and saturation the plate current will increase with the plate potential in some such manner as is indicated by the curve of Fig. 132, in which the ordinates represent plate current and the abscissæ plate potentials. This is called a "plate current—plate potential" characteristic curve.

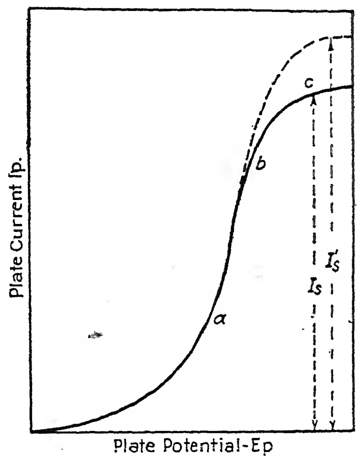


FIG. 132.—Plate current—plate potential characteristic.

Proceeding from the lower current and voltage values, it is typical of such curves to bend decidedly upward at first until some such point as a is reached, then gradually to become more nearly straight, as between a and b, and, finally, to bend rather sharply in the reverse direction until it becomes nearly horizontal. The characteristic, therefore, indicates that at first the current rises more rapidly than the potential, then, as the curve straightens, the action is more nearly linear, the current increasing in direct proportion to the potential. Beyond the linear range the knee in the curve indicates that further increases in plate poten-

tial cause less and less increase in current. At about the point c the curve becomes nearly horizontal so that further changes in potential produce little effect on the current. Obviously, this point c is about the saturation point, and we may refer to the corresponding saturation current as I_s.

If it is desired to secure from this tube higher plate current than I_s, it can be done only by increasing the filament temperature so as to cause greater emission. Under this new temperature condition the upper part of the curve of Fig. 132 would assume such form as that indicated by the dashed curve, a new saturation point c' being reached, with a corresponding saturation current I_s'.

Similarly, Fig. 133, copied from van der Bijl, shows three current-voltage characteristics for the same tube taken at three

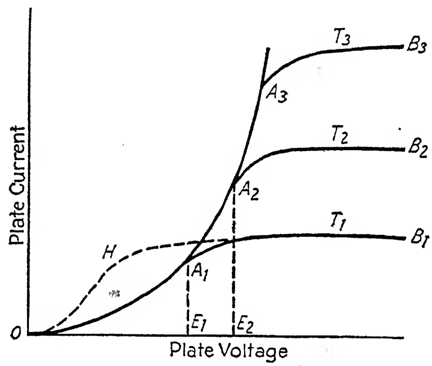


FIG. 133.—Plate current—plate voltage characteristics for different temperatures.

different temperatures T_1, T_2 and T_3, respectively. Here we have three successive values of the saturation current, corresponding to the three different temperatures. It is to be noted that the lower part of all three curves are common, showing that the current-voltage characteristic below the lowest saturation point is independent of temperature. The dotted curve H of Fig. 133 shows the form of curve that would result for temperature T_1, if the electrodes were moved closer together. As might be expected, below saturation larger current would flow for any applied voltage, but above the saturation point the current would remain the same.

Clearly, for those purposes where the tube is required to vary its space current in response to changes in applied voltage, the useful operating ranges are those represented by the lower portions of these curves, OA_1, OA_2 or OA_3 for temperatures T_1, T_2

or T_3, respectively. Beyond these ranges the current is practically independent of the applied voltages. The condition of saturation represented by the horizontal portions of the curves of Figs. 132 and 133 is sometimes called "voltage saturation." This term implies the condition for any given temperature, where the current is at its maximum, regardless of any higher voltage that might be applied.

There is another kind of saturation, however, to which we may now refer in connection with the "current-temperature characteristic curve" of a tube. The curves of Figs. 129, 130 and 131 showed the total emission (saturation current) for various temperatures of the filament, or for various amounts of energy applied to heat it. These curves were plotted on the assumption that *the plate was always held at high enough potential to attract all of the emitted electrons to it.* If now we hold the plate at some lower potential E_1, not high enough to attract all the emission at the higher temperatures, a different state of affairs will result. The curves showing the rise of current with increasing temperature will not continue to rise indefinitely, as in those figures, but at some rather definite temperature will flatten out, indicating that the current is independent of temperature for all higher temperatures. This is called "temperature saturation," and describes the condition, for any given plate voltage, where the current is at its maximum, regardless of any higher temperature that might be applied.

Figure 134, taken from van der Bijl, shows typical current-temperature characteristics of a tube for three different plate potentials E_1, E_2 and E_3. With the plate held at the constant potential E_1, some such curve as OC_1D_1 will represent the plate-current values as the temperature rises. The flattening of the curve beyond the point C_1 means that, for that particular plate potential E_1, the current will not increase no matter how much higher the temperature may become. This is evidently temperature saturation for that particular plate potential.

Again, for other constant plate potentials, such as E_2 and E_3, higher temperature saturation points are reached at C_2 and C_3, beyond which no further increase of filament heat will serve to increase the current.

In the case of Fig. 133 it is the total emission from the filament at any temperature that limits the current for higher voltages. In the case of Fig. 134 the cause for the limitation in the current

flow at higher temperatures is not quite so apparent. It is mainly the effect of the space charge in repelling the escaping electrons. As the filament temperature rises the swarm of electrons in the space between the electrodes increases in density. The negative space charge thus increases in intensity until finally its repelling effect on the electrons emerging from the

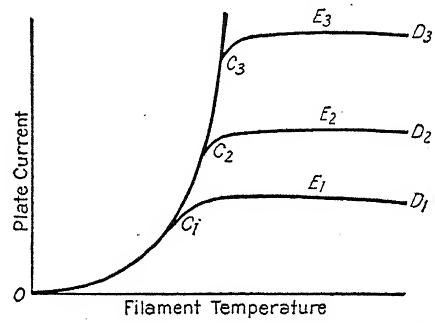


Fig. 134.—Plate current—filament temperature characteristics for different plate voltages.

filament counteracts the attraction by the positive plate. Beyond this point the current does not increase with further rise in temperature and consequent emission.

The curves of Figs. 133 and 134, while similar in form, have quite different meanings. They are complementary and together tell the general story of the current-voltage temperature charac-

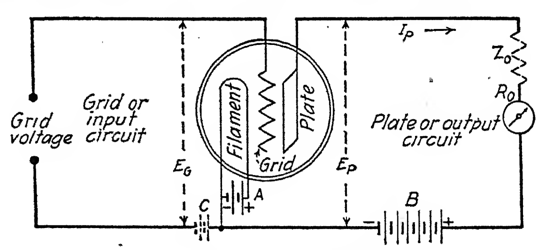


Fig. 135.—Action of three-electrode tube.

teristics of the two-electrode tube. It will be observed that the operating range in which the current varies with the voltage is represented by the stem of the curve in Fig. 133 and by the horizontal branches of the curve of Fig. 134. On the other hand, the range of saturation, where the current is independent of voltage, is represented by the horizontal branches of Fig. 133 and the stem of Fig. 134. The plate or anode currents corre-

sponding to the portion OC_1 of Fig. 134 are evidently saturation currents, for they are the same for all plate voltages, E_1, E_2, E_3 and higher.

The three-electrode tube, or audion (Fig. 135), has three possible independent variables: filament temperature, plate voltage and grid voltage. Ordinarily, in operation the filament temperature, once brought to the desired degree by adjusting the filament current from battery A, is held constant. Also in operation the battery B, which furnishes the plate potential, ordinarily maintains constant voltage. Practically, therefore, for any given adjustment of the filament temperature and plate potential, the grid potential E_g remains as the independent variable. The important thing to study, therefore, is the effect of changes in grid potential on the plate current.

To obtain this effect is the whole sum and substance of the purpose of the three-electrode vacuum tube. We may illustrate it in connection with the concrete example of Fig. 135, which may be taken as typical. The potential of the grid E_g, by which we mean its potential with respect to the filament, is made susceptible to influence by the incoming voice, signal or other currents which it is desired to detect, amplify or otherwise modify. In other words, the varying currents from an incoming telephone line, or from a radio antenna circuit, or from any other source of "input" are so associated with the input or grid circuit of the tube as to cause corresponding changes in voltage across the terminals marked "grid voltage." The potential of the grid is thus made to vary in accordance with the current variations of the input. These changes in grid potential, often exceedingly minute, operate as a control on the very much larger current flowing from the battery B through the plate circuit, which is the output circuit of the tube. Here a coil Z_0, symbolizing the impedance of the load, or whatever device is to consume the output, is shown in the output circuit, together with an instrument for measuring the output current.

In this connection it is to be remembered that the potential of the filament varies along its length due to its IR drop. The filament, therefore, has no definite single potential, and, in order to have a fixed point of reference, it has become customary, in speaking of grid potential, to refer it to the negative end of the filament. The same point of reference is usually employed with respect to the plate potential. Sometimes, however, the

circuit arrangements are such as to make it more convenient to refer the grid and plate potentials to the positive side of the filament. This choice of a reference point within the range of potential along the filament may or may not be of considerable moment, according to the relative potentials employed. In the Edison effect (Fig. 121) for example, it was of great importance, for whether the effect occurred or not depended on which end of the filament was chosen for connection with the plate.

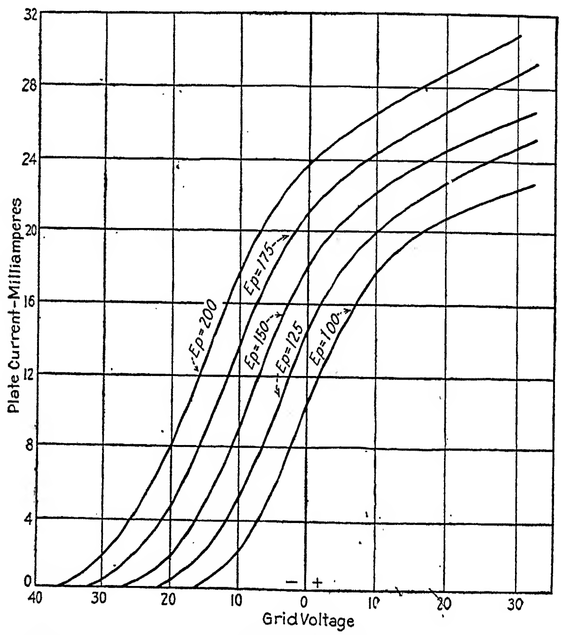


Fig. 136.—Plate current—grid voltage characteristics for different plate voltages.

Returning now to the relationship between grid voltage and plate current, we will first take the case where the load Z_0 is not present, and there is no considerable resistance R_0 in the output circuit external to the tube; in other words, where the external impedance and resistance are zero. We may assume that in Fig. 135 the filament temperature has been adjusted to secure a proper emission and that the plate voltage E_p is held constant at some definite initial value, say at 100 volts, by the battery B. Changes are to be made in the applied grid voltage and their effects on the plate current observed and measured by the

instrument in the output circuit. Figure 136 shows the effect on the plate current of such changes in grid voltage, for several different values of plate voltage.

An examination of the curves of Fig. 136 shows that small changes in grid potential will be as effective as larger changes in plate potential in producing variations in plate current. We see from the $E_p = 100$ curve, for instance, that by changing the grid potential from -10 to 0 volts, a change in plate current from 2 to 10 milliamperes results, or 0.8 milliampere per volt. On the other hand, holding the grid potential constant at zero, for instance, and raising the plate potential from 100 to 125 volts causes the plate current to increase from 10 to 14 milliamperes. It takes, therefore, a change of 25 volts in plate potential to effect a change of 4 milliamperes in plate current, about 0.16 milliamperes per volt. For this particular tube, therefore, changes in grid potential are about five times as effective in varying the plate current as are changes in plate potential. This ratio between the changes in plate and grid voltages required to produce a given change in plate current (in this case $0.80/0.16 = 5$) is called the "amplification constant" of the tube. It is one of the most important characteristics of the vacuum tube. In mathematical treatises it is usually symbolized by the Greek letter μ. The relationship is expressed by the equation $\mu = \Delta e_p/\Delta e_g$, in which Δe_p and Δe_g are, respectively, the changes in plate and grid potentials necessary to produce the same change in plate current.

It follows from what has just been said that a change in grid potential produces μ times as great a change in plate current as an equal change in plate voltage. The practical significance of this is that a variation in the grid potential produces the same change in the plate current as introducing an electromotive force into the plate curcuit μ times as great as the grid potential variation.

The amplification constant is not strictly a constant for a given tube, but, within the range of plate and grid voltages at which tubes are ordinarily operated, it is sufficiently so for practical purposes. It depends solely on certain of the mechanical dimensions of the tube, such as the distance between the grid and plate, the mesh of the grid and the diameter of the grid wire.

Van der Bijl developed empirically a simple formula relating the amplification constant to the pertinent structural dimensions of the tube. It is:

$$\mu = Cprn^2 + 1,$$

in which p (Fig. 137) is the distance between plate and grid, r the diameter of the grid wires, n the number of grid wires per unit length, and C a constant the value of which depends on the general type of tube. For tubes of the type shown in Fig. 124, in which the filament, grid, and plate are in parallel planes, the value of C is about 80. Strangely enough (see Fig. 137) the distances f from filament to grid and $f + p$ from filament to plate, do not appear in the equation; indicating, as is the fact, that the amplification constant μ is independent of these two dimensions.

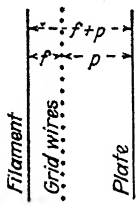


FIG. 137.—Distances between tube elements.

Another important characteristic of the three electrode vacuum tube is the plate resistance R_p. This is the resistance of the plate curcuit when there is no resistance in the external path; that is, it is the resistance of the internal path from filament to plate. While this internal characteristic of the tube is often referred to as the "tube impedance," or as the "plate impedance," it is of the nature of a pure resistance, except where very high frequencies are involved. This resistance is due to the unrecoverable work done on the electrons. In moving from filament to plate they acquire velocity, principally because of the attraction of the positive plate, and, having velocity, they also have kinetic energy. On striking the plate the kinetic energy is transferred to the molecules of the plate taking the form of thermal agitation, which manifests itself in heat. There is no difference in phase between the electromotive force and current involved in this work, and so the work becomes merely

$$W = I_p^2 R_p = E_p I_p$$

where E_p and I_p are respectively the total direct-current plate voltage and plate current. The direct current resistance of the filament-to-plate circuit within the tube then becomes:

$$R_p = \frac{E_p}{I_p}.$$

The alternating-current resistance r_p, that is the internal resistance of the plate circuit to an alternating component

impressed on the direct current in that circuit, will be the ratio of the voltage increment ΔE_p to the resulting current increment ΔI_p, so that the alternating current resistance over any small range affected is

$$r_p = \frac{\Delta E_p}{\Delta I_p}.$$

This pure resistance aspect of the plate impedance, whether for direct or alternating currents, is a sufficiently close approximation for most tubes for frequencies up to several hundred thousand cycles per second. For higher frequencies the electrostatic capacities between the tube electrodes begin to introduce appreciable wattless currents; and, therefore, capacity reactance must be taken into account together with the resistance.

Evidently the ratio of the direct-current values E_p/I_p depends on the slope of the straight line connecting the corresponding point on the current-voltage characteristic and the origin, while the ratio of the differentials $\Delta E_p/\Delta I_p$ depends on the slope of the small section of the characteristic involved. These slopes cannot be equal except at two points on the compound curve of the characteristic, one at the origin and the other well up on the knee of the curve. Over the intermediate parts of the characteristic, where the curve is concave upward, a little thought will show that the ratio of the alternating differentials will be less than that of the direct-current totals, and that, therefore, the alternating-current resistance of the tube must always be less than the direct-current resistance throughout the useful range of working. Van der Bijl shows, on page 195 of his book, that, within the range most useful for amplification, that is, along the nearly straight portion of the characteristic, a fair estimate of the alternating-current plate resistance, when the grid voltage is held at zero, is obtained by taking it as half the direct-current resistance.

The determination of the direct-current plate resistance R_p from such curves as those of Fig. 136 is obvious. For instance, at zero grid voltage, a plate voltage E_p of 100 gives a plate current of about 10 milliamperes, with a resulting direct-current plate resistance R_p of about 10,000 ohms, and an alternating-current plate resistance r_p of about 5,000 ohms. A few trials on various parts of the diagram will show that the plate resistance is not constant for the tube represented by these curves but varies with different values of the grid and plate potentials. For this reason,

in expressing the tube resistance, it is customary to refer it to zero grid potential and to some certain plate potential.

Unlike the amplification constant μ, the plate resistances R_p and r_p are not principally dependent on the geometry of the tube. They depend in part on the structure of the tube, and also on the values of the plate and grid voltages at which the tube is operating. As to the structural dimensions affecting plate resistances, the plate resistances increase with the length of path, *i.e.*, with the distance $f + p$ (Fig. 137); and they decrease as the areas of plate and filament increase. Owing to their variation with plate and grid voltage, the plate resistances may best be specified for a given tube as functions of the plate voltage, holding the grid voltage constant at zero. Under this condition of zero grid voltage the plate resistance r_p varies with the plate voltage in some such manner as shown in Fig. 138.

It will be realized that the variation in the plate resistance of a tube with the varying plate voltages is due to the non-linear characteristics of the tube, which means simply that the current changes are not proportional to voltage changes. If the tube action were linear, then the current-voltage characteristics would be straight instead of curved lines. I_p would then vary in proportion to variations in E_p, and the resistance and conductivity of the path from filament to plate would each be constant.

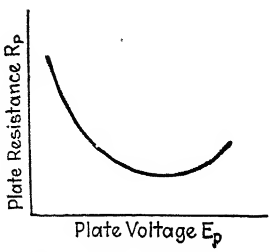


FIG. 138.—Variation of plate resistance with plate voltage.

It is often important to secure as nearly linear tube response as possible, as, for instance, when amplification without distortion is desired. Some curvature in tube characteristic is unavoidable, but, as will be shown, by choosing the straightest part of the characteristic on which to work and by properly designing the portion of the output circuit external to the tube, a near enough approach to strict linearity may be had to secure practically distortionless amplification.

It must be kept in mind that the characteristics of a tube, *per se* may be quite different from the characteristics of the tube with its external circuit. The curves of Fig. 136, for instance, were taken with almost no resistance in the plate circuit external to the tube. As a result, the plate battery B being of constant

voltage, the potential E_p (Fig. 135) from plate to filament remained constant.

Another set of characteristics results when there is a considerable resistance R_o in the external plate circuit. In this case, the actual voltage across the plate-filament space will be modified by the IR drop through the external circuit. Inspection of Fig. 135 will show that with the resistance R_o present in the output circuit, the plate potential E_p will not be constant and equal to

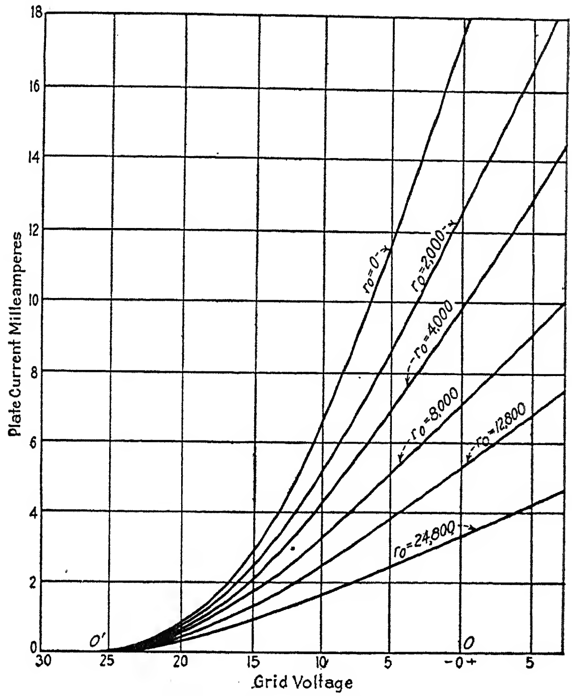


FIG. 139.—Plate current—grid voltage characteristics for different output resistances.

the battery voltage but will vary and be less than the battery voltage because of the voltage drop for different current values through the external resistance R_o.

Figure 139, taken from van der Bijl, gives a set of grid voltage-plate current characteristics for the same tube with different resistance values in the output circuit external to the tube. In taking these the voltage of the battery B (Fig. 135) was held constant, as before, but the voltage from filament to

plate was, for all except the $R_o = O$ curve, less and varying, being, obviously,

$$E_p = E_b - R_o I_p.$$

Two principal effects will be noticed as a result of the introduction of the external resistance: First, a given change in grid potential causes smaller and smaller changes in plate current as the external resistance in the plate circuit is increased. This is indicated by the fact that the slope of the curves decreases with the higher values of R_o. Second, the changes in plate current become more nearly proportional to the grid voltage variations as the value of the external resistance increases. The

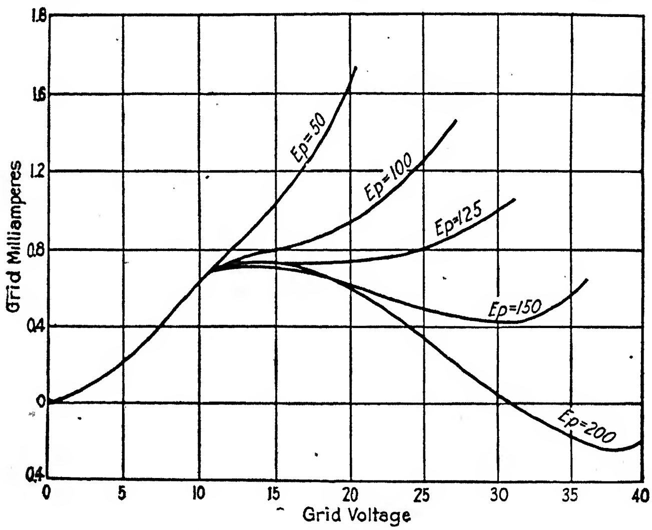


FIG. 140.—Grid current—grid voltage characteristics for different plate voltages.

action becomes practically linear over a considerable range of grid potential for the higher external resistances, the curves becoming more nearly straight as the value of R_o is increased. It was pointed out by Dr. H. D. Arnold that this substantially straight-line characteristic was secured when the external resistance in the output circuit was equal to or greater than the plate resistance of the tube itself. This is an important point where distortionless amplification is sought.

Under certain circumstances electrons flow from the filament to the grid. This flow is spoken of as "grid current." Obviously, when the grid is negative to the filament there can be no

such current for exactly the same reason that there can be no plate current when the plate is negative to the filament. The only range of grid potential, then, that is it necessary to explore in investigating grid current, is that in which the grid is positive to the filament.

The diagram of Fig. 140, shows grid-current values for various positive grid potentials, as given by van der Bijl. The various branches of the curve represent the results secured with the plate held at the respective potentials indicated. Of course, this grid current does not represent the entire electron stream from the filament but only that part which is caught by the grid. The bulk of it, under ordinary circumstances, goes on to the plate.

The falling off of the grid current after reaching a maximum, as indicated in the curves of higher plate potential, is due to secondary electron emission from the grid. Secondary emission has already been referred to, and, in this case, it is caused by

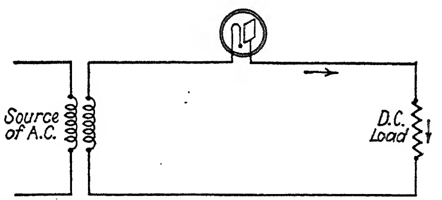


FIG. 141.—Rectification with single tube.

the grid itself becoming a source of emission under the impact of the electrons coming from the filament. In other words, some of the electrons from the filament strike the grid with such force as to blast off some of its own electrons. The grid itself thus becomes a source of emission setting up counter currents which tend to reduce the net grid current.

A glance at Fig. 135, will show that any flow of grid current must take place in the input circuit of the tube. Its effect, therefore, will be added to the legitimate input of the tube and may serve to distort it, and thus, also, distort the output. The way to prevent this, in cases where this factor is of importance, is to keep the normal grid voltage sufficiently negative to assure its remaining so even when subject to the influence of the varying input voltage. For this purpose it is customary to add a small battery of a few volts in the grid circuit, as indicated by the dotted battery in Fig. 135. This may be associated with the

grid circuit in a variety of ways, but always with the purpose of keeping the grid slightly negative to the filament. It is commonly called the "grid bias battery" and on diagrams of vacuum-tube circuits it is usually referred to by the letter C.

Rectification.—It is apparent that the Fleming valve, or the two-electrode thermionic vacuum tube, is by its very nature a current rectifier. It conducts only in one direction. Figure 141 shows a simple rectification circuit. The source of alternating

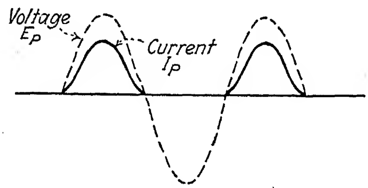


FIG. 142.—Rectified current from single tube.

current to be rectified is associated with the plate circuit by means of the input transformer at the left. The load, or device receiving the rectified current, is indicated at the right. The battery for heating the filament is omitted from the diagram for simplicity.

Once during each cycle, in such a circuit, the plate will be rendered alternately positive and negative with respect to the filament. During each half cycle, when the plate is positive,

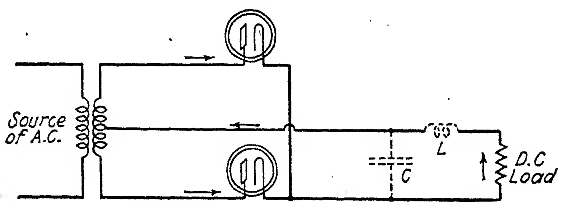


FIG. 143.—Rectification with two tubes.

current will flow, and, during each alternate half cycle, when the plate is negative the current will be blocked. The current which actually flows through the plate circuit will be a series of unidirectional pulsations, as indicated in Fig. 142.

For many reasons such a current as that of Fig. 142 would be highly objectionable. Such pulsating current would introduce serious noises in the receivers of telephone systems if the frequency of the pulsations (or of any of the harmonics) was within the audible range.

By using two tubes connected as shown in Fig. 143, both half cycles of current are used. Obviously, in this arrangement, when the plate of one tube is negative that of the other is positive. Therefore, the half wave which is blocked by either tube will be transmitted by the other. As will be seen, the transmitted impulses from each tube pass through the load in the same direction but in alternate half cycles. The current flow through the load therefore becomes something like that of Fig. 144.

While this condition would represent an improvement over that of Fig. 142, as regards smoothness of current, it would still be far from satisfactory. Connecting a large condenser C across the load, as shown in dotted lines in Fig. 143, would go far toward improving the smoothness of the current, the condenser acting as a reservoir to receive current during the crests of the voltage waves and to supply current during the troughs. A still further smoothing effect may be gained by adding the impedance coil L, in series with the load, as in dotted lines in Fig. 143. This stores energy in a magnetic field when the current is increasing, at the same time holding the current back, and gives up this energy when the current is decreasing, tending to maintain the current at its maximum value. If such a combination of properly proportioned capacitance and inductance does not give a sufficiently smoothed-out current, almost any desired degree of smoothness may be obtained by properly designed networks of condensers and reactance coils, forming filters, which may be specifically pointed to the elimination of whatever frequencies remain as ripples on the continuous current.

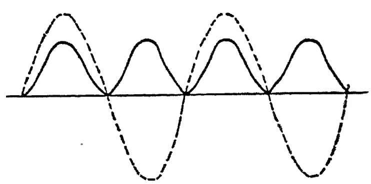


FIG. 144.—Rectified current from two tubes.

Amplification.—Unlike the induction coil or the alternating-current transformer, which can amplify the primary voltage but not the power, the vacuum tube amplifier can be used to amplify both the voltage and the power of the input. Herein lies the greatest usefulness of the thermionic vacuum tube. Of course, the tube does not create new power, but, under the control of the feeble input, it releases the power of a battery or other source in such manner that there is a corresponding output very

much magnified as to voltage or power, or both. When it is important to do so, as in the amplification of speech currents, the output may be made an almost exact replica of the input, but on a magnified scale. This means that there may be amplification without distortion. In some cases we are concerned merely with the amplification of voltage, in others in the amplification of power.

Voltage amplification is measured as the ratio of the output to the input voltages. Where the input is alternating, as in the case of voice currents received over a long telephone line, the voltage amplification will be the ratio of the voltage of the alternating component in the output circuit to the voltage of the alternating input.

We have already seen that the amplification constant μ of a tube is the maximum voltage amplification of which the tube is capable. It is a characteristic of the tube itself and is independent of the related circuits. Under ordinary circumstances a smaller amplification than the μ of a tube is realized from the combination of the tube and its external circuits and associated apparatus, and we may inquire how the voltage amplification actually attained is modified by the external circuit characteristics.

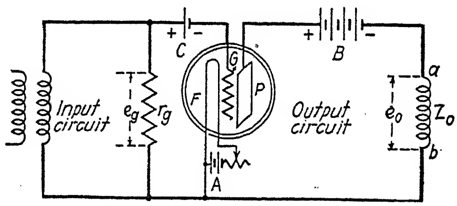


FIG. 145.—Amplification.

Consider such a circuit as that of Fig. 145. If the alternating voltage impressed on the grid is e_g, then, as has been stated, the alternating voltage e_p acting in the plate circuit will be μ times e_g; so that

$$e_p = \mu e_g.$$

This alternating voltage e_p is the active alternating voltage in the plate circuit and, therefore, the alternating current i_p in the plate or output circuit will be

$$i_p = \frac{\mu e_g}{R_p + Z_0},$$

where R_p is the internal plate-to-filament resistance of the tube,

and Z_0 is the total resistance or impedance of the external part of the output circuit.

The available alternating voltage outside the tube will be the electromotive force e_0 across the terminals of the external resistance or impedance Z_0. Evidently then,

$$e_0 = i_p Z_0,$$

and, substituting the value of i_p derived in the last paragraph, we have

$$e_0 = \frac{\mu e_g Z_0}{R_p + Z_0}.$$

Designating the actual voltage amplification by μ', and remembering that it is the ratio of output to input voltages, we have

$$\mu' = \frac{e_0}{e_g} = \frac{\mu Z_0}{R_p + Z_0}.$$

From this it is evident that the voltage amplification μ' will always be somewhat less than the amplification constant μ of the tube, and will become equal to it only when the external impedance is infinitely large with respect to the plate resistance of the tube. Thus, in the limiting case when the output circuit is open, $\mu' = \mu$, which means that the actual voltage amplification is then equal to the amplification constant of the tube.

Since μ is evidently the limiting factor in voltage amplification, it is desirable that tubes intended for voltage amplification should be designed for as high amplification constant as possible. This is purely a matter of internal structural dimensions. A form of Western Electric tube largely used for voltage amplification has an amplification constant of 40. It is possible, however, to design tubes in which the values of μ are several hundred.

As shown by the equation just derived, the actual voltage amplification μ' depends not alone on the μ of the tube but on the impedance characteristics of the output circuit as well. The equation shows that in order for the realized voltage amplification μ' to become a fairly large proportion of the maximum μ possible with the tube, the external impedance Z_0 must be several times the plate resistance R_p. We must distinguish, however, between the cases where the external impedance is non-inductive and those where it is inductive.

Taking first the case where the external impedance is a pure resistance R_0, the equation for μ' becomes

$$\mu' = \frac{\mu R_0}{R_p + R_0}.$$

Assuming a tube with a μ of 10, we see by substituting different values of R_p and R_0 that, when the external resistance is just equal to the internal, the resulting amplification μ' is only 5. In order to make μ' as great as 90 per cent, for instance, of μ, it is necessary to increase R_0 to 9 times the value of R_p.

On the other hand, taking, as an extreme case, one where the external impedance is pure reactance X_0 with negligible resistance, we can no longer determine the total impedance of the output circuit by the direct addition of the internal and external components, but must add them vectorially. The equation thus becomes

$$\mu' = \frac{\mu X_0}{\sqrt{R_p^2 + X_0^2}}.$$

Again substituting trial values of R_p and X_0 in this equation, we see that for a tube having a μ of 10, the actual amplification μ' becomes about 7 when the external reactive impedance is just equal to the plate resistance, and reaches a value of about 90 per cent of μ when the external impedance is only twice the internal.

We see, then, that except in cases where non-inductive plate circuits are especially required, it is advantageous to make the external impedance Z_0 largely reactance instead of pure resistance.

Another advantage of using reactance instead of resistance in the external circuit is that, on account of its lower direct-current resistance, not so much of the plate circuit battery voltage is lost in the resistance drop through it. Thus in Fig. 145 if the coil Z_0 is made a highly reactive choke coil with low ohmic resistance, the required highly amplified alternating voltage across its terminals may be obtained with only a small IR drop of plate potential through it.

On the other hand, the disadvantage of employing reactance instead of pure resistance in the output circuit lies in the fact that the impedance of inductive circuits varies widely for different frequencies, whereas non-inductive circuits offer the same impedance to all frequencies. The tendency to distortion of voice currents is thus largely avoided by employing non-inductive circuits.

Power amplification is measured as the ratio of the output power to the input power. Because the output power is of considerable magnitude while the input power is often extremely small, the ratio of the two, expressing the degree of power amplification, may be very large. Van der Bijl states that, with a specially designed tube, he has secured a power amplification of 3,000 fold.

We have seen in Chap. VIII that the power of an alternating current is represented by the product of the current and the electromotive force when the two are in phase with each other. When, however, there is a difference in phase, this product must be multiplied by a third factor, the cosine of the angle of phase difference. We have, then,

$$P = EI, \text{ for non-inductive circuits,}$$

and

$$P = EI \cos \theta, \text{ for inductive circuits,}$$

where

θ is the angle of lag or lead.

Let us consider the output power P_0, that is, the power expended in Z_0 (Fig. 145). We have already developed the equations for the electromotive force and the current in the plate circuit, these being

$$e_0 = \frac{\mu e_g Z_0}{R_p + Z_0} \text{ and } i_p = \frac{\mu e_g}{R_p + Z_0}.$$

From these the general equation for output power is

$$P_0 = e_0 i_p \cos \theta = \frac{\mu^2 e_g{}^2 Z_0}{(R_p + Z_0)^2} \cos \theta,$$

from which it is seen that the output power increases as the square of the amplification constant and of the grid voltage, directly as the cosine of the angle of lag introduced by the reactance of the external circuit, and is also dependent in a somewhat more complex way on the relationship between the external impedance and the plate resistance. When the output circuit is non-inductive, the angle θ becomes zero and the power factor, $\cos \theta$, unity. In that case the amplification is the same for all frequencies.

The input power is not so easily dealt with if we consider only the power actually consumed by the grid alone. Except for very high frequencies, several hundred thousand cycles per second and upward, the impedance of the grid-filament path

within the tube is almost infinite. The power consumed in this path is, therefore, practically zero, and, as a result, the ratio of output to input power P_0/P_g is very great and indeterminate.

As a matter of practical operation, however, it is usually desirable for purposes of stability or for other reasons to shunt the grid circuit by a very high pure resistance in r_g (Fig. 145). The power consumed in this resistance must be furnished by the input circuit and may, in many cases, constitute the major part of the input power. In such a case we will not be in serious error if we assume this to be the total input power for purposes of calculating power amplification. Thus. the input power becomes

$$P_g = e_g i_g = \frac{e_g^2}{r_g}.$$

From these two equations for output and input power, we derive the power amplification equation

$$\eta = \frac{P_0}{P_g} = \frac{\mu^2 e_g^2 Z_0 \cos\theta/(R_p + Z_0)^2}{e_g^2/r_g} = \frac{\mu^2 r_g Z_0}{(R_p + Z_0)^2}\cos\theta.$$

For a non-inductive output circuit, Z_0 becomes R_0, and the power factor unity. The power amplification is, then,

$$\eta = \frac{\mu^2 r_g R_0}{(R_p + R_0)^2},$$

which means that for non-inductive input and output circuits the power amplification is independent of the frequency, varies as the square of the amplification constant of the tube, directly as the grid resistance, and is also dependent on the relationship between the internal and external resistances of the plate circuit.

It can readily be shown that for output circuits containing reactance as well as resistance, the power amplification is a maximum when $R_p = Z_0$; that is, when the tube works into a circuit having an impedance numerically equal to its own plate resistance. In the special case of a non-inductive output circuit, maximum power amplification is attained when $R_p = R_0$; that is, when the tube works into a non-inductive circuit having an ohmic resistance equal to its own plate resistance. This relation between internal and external impedance for maximum power amplification plays an important part in the design of vacuum-tube circuits.

As a result of this elementary discussion of voltage and power amplifications, it may be stated, in a general way, that voltage

amplification can never exceed the amplification constant μ of the tube and approaches it more closely as the ratio of the external impedance or resistance of the output circuit to the plate resistance of the tube increases. To realize a fair proportion of the voltage amplification of which a tube is capable, the impedance of the output circuit should be at least several times the plate resistance.

Power amplification, on the other hand, may be indefinitely greater than the amplification constant of the tube. It reaches a maximum value, for both inductive and non-inductive circuits, when the output impedance or resistance is numerically equal to the plate resistance of the tube.

We may now consider the conditions for distortionless amplification—vital in speech transmission though relatively unimportant in some other applications. In Fig. 146 is shown, by heavy

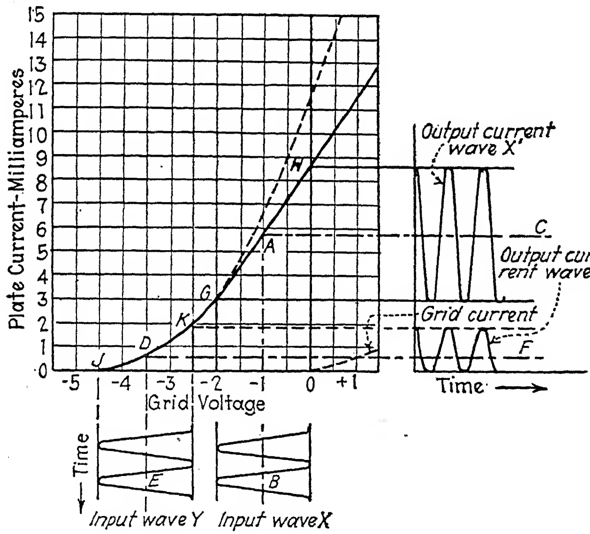


FIG. 146.—Effect of curvature in characteristic on distortion.

line, a plate current-grid voltage characteristic. We may consider this curve to represent the action of a tube with a high impedance in the output circuit, this being indicated by the closeness with which its upper portion approximates a straight line. The same tube with zero impedance in its external circuit would show a characteristic such as given by the dashed curve.

The range of grid voltage covered by this curve is from $-4\frac{1}{2}$ to $+1\frac{1}{2}$. For grid voltages below $-4\frac{1}{2}$ there is no plate current.

When the grid voltage changes from negative to positive a small grid current appears, increasing as the grid potential becomes higher. This is indicated in the lower right-hand corner of Fig. 146.

If we give the grid a "normal" potential of -1 volt as by the grid-bias battery C of Fig. 145, then it is obvious that any alternating input wave will alternately raise and lower the grid potential about this "normal" potential. Thus, in Fig. 146 the sinusoidal input wave X having a maximum amplitude of 1 volt will cause variations of the grid potential from 0 to -2 volts. For these maximum grid variations the plate current will vary along the portion G to H of the curve, indicating changes in plate current between limits of about 3 and 8.6 milliamperes. The resulting plate-current wave is shown at X' at the right of the figure. The choice of a normal grid voltage of -1 determined the axis AB about which the incoming wave X would vary, and this, in turn, determined the axis AC representing a plate-current value of 5.8, about which the output current wave X' would vary. Evidently each increment or decrement of the input voltage will produce an exactly proportional effect on the output-current value, since the portion of the curve G-H along which the variations occur is a straight line. This means distortionless amplification.

If now, using the same normal grid voltage, the input wave increased so as to have an amplitude greater than 1 volt, it is evident that each positive maximum would cause the grid to become momentarily positive. During these instants the grid would take current, and these pulses of grid current, since they occurred in the input circuit, would distort the input wave and, as a final consequence, the output wave.

Assume that we give the grid a normal bias of $-3\frac{1}{2}$ volts and apply an input wave Y having an amplitude of 1 volt. The grid variations will now occur about the axis DE and between values of $-2\frac{1}{2}$ and $-4\frac{1}{2}$ volts. The resulting plate current wave Y' will vary about an axis DF, with a current value of about 0.6 milliampere and between limits indicated by the points J and K on the curve, having, respectively, current values of 0 and 1.8 milliamperes. We are now working on the curved portion J-K of the characteristic and the action is no longer linear. Evidently a change of 1 volt on the grid below $-3\frac{1}{2}$ will produce a very much smaller change in plate current than a 1-

volt change in the other direction. The unsymmetrical form of the output-current curve Y' due to this distortion is obvious.

The principal requirements for distortionless amplification, so far as the tube characteristic is concerned, are:

1. That a sufficiently high impedance be placed in the external circuit to give a straight-line characteristic over a proper working range.

2. That a normal grid voltage be adopted such that the maximum potential changes caused by the input wave will lie within the straight-line portion of the characteristic and will never cause the grid to become positive.

Frequently, more amplification is required than can be secured through a single tube. In this case "multistage amplification" may be employed. The tubes are arranged "in cascade," the output of one tube forming the input of the next, and so on.

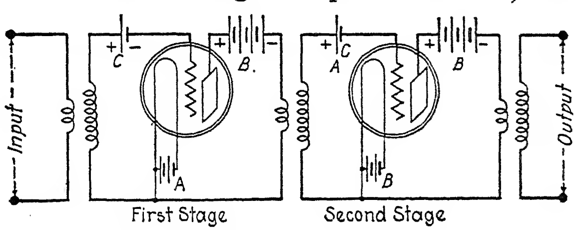


FIG. 147.—Two-stage amplification.

Such an arrangement is shown, stripped of detail, in Fig. 147. The action will be sufficiently clear from what has already been stated, but the particular method, here shown, of coupling each stage and its adjoining circuits deserves some attention.

This is a transformer-coupled circuit—a type largely used where non-inductive circuits are not imperative. Its disadvantage lies in the fact that the impedance of the transformers will vary somewhat for different frequencies and thus introduce distortion. Its advantages lie principally in the facts that a voltage step-up may be had in each of the interstage transformers, and that by properly designing the transformers each tube may be made to work into a proper impedance to give the best working efficiency. King says that when uniform amplification over a wide band of frequencies is not required, the interstage transformer may be designed to step up the voltage as many as thirty or forty times.

Other types of amplifier coupling will be referred to in other parts of this work dealing with the various applications of vacuum-tube amplifiers.

Oscillation.—The fact that the vacuum tube is inherently an amplifier leads to another of its important uses. It may be used as a generator of oscillatory currents, or, more properly, as a converter of continuous currents into alternating currents. As such, since its range of frequency is practically unlimited, it forms a valuable tool in the hands of the communication engineer. In wire telephony the principal use of the vacuum-tube oscillator is in the generation of the high-frequency carrier waves used in carrier-current systems of multiplex transmission. In radio, or wireless telephony, it has a similar use—the generation of the high-frequency carrier wave which is sent out from the transmitting station. In each of these cases it is the continuing chain of inaudible carrier waves which is modified in amplitude by the telephone transmitter so that its variations are in accordance with the form of the much lower frequency sound waves that are to be reproduced.

The way in which continuous and self-sustaining oscillations may be set up in vacuum-tube circuits is not difficult to understand. It is merely a case of one part of a closed system reacting on another part, which in turn reacts again on the first part, and so on. A somewhat analogous action is that of an ordinary vibrating battery bell or buzzer. Another is the singing or howling which sometimes occurs when the receiver is held close to the mouthpiece of the transmitter of the same telephone. Briefly, in the vacuum-tube oscillator the output circuit carries more energy than is supplied to the input circuit. If a part of this output energy is fed back to the input circuit as by an inductive or other type of coupling between the two circuits, and if this feed-back occurs at just the proper time intervals to secure the proper phase relations, the energy so received as input may act again in the ordinary way on the output, which, in turn, will react through the coupling on the input, and so on. The time relationships are dependent on the natural periods of the circuits, as determined by their resistance, capacitance and inductance, and on the manner in which the input and output circuits are connected or otherwise associated with each other.

There are many different oscillation circuits, but a simple one is shown in Fig. 148. Associated with the plate or output circuit

is an oscillatory circuit containing resistances r_1 and r_2, an inductance L_2 and an adjustable condenser C. This constitutes a resonant circuit, and it may be tuned to a desired frequency by adjusting the capacity of C. The input circuit contains another inductance, L_1, so related to L_2 that there will be mutual inductance M between them. The mutual inductance between L_1 and L_2 constitutes the inductive coupling between the output and the input of the tube.

Let us assume a steady state to be established in the circuit. Suppose the current in the plate circuit to decrease slightly, due, say, to a fluctuation in battery voltage. This decrease will allow a release of energy due to partial collapse of the magnetic field of L_2. This released energy will drive a current around the oscillatory circuit charging the condenser C. The condenser will discharge back through the inductance and resistances, again increasing the magnetic field, which will again decrease and charge the condenser, and so on, each surge being less intense because of loss of energy in the resistances. The oscillations will continue, dying out more or less rapidly according to whether the resistances are large or small. If, as is the case, the changing magnetic field of L_2 cuts the turns of L_1, an electromotive force is induced in the grid circuit. This electromotive force, with its multiplied effect due to the amplification of the tube, will accentuate or suppress the disturbance in the plate circuit, depending on whether L_1 is connected in such a way as to put the grid voltage in phase or 180 degrees out of phase with the alternating-current plate component. If it is so connected that these voltages are in phase, the oscillations in the circuit $L_2Cr_1r_2$ will be prolonged. Further, if the mutual inductance between L_1 and L_2 is great enough, or the resistances r_1 and r_2 are small enough, or the voltage amplification constant of the tube is great enough, or a favorable combination of these factors exists, the oscillations will be prolonged indefinitely. If the resistances are less, or the amplification constant, or mutual

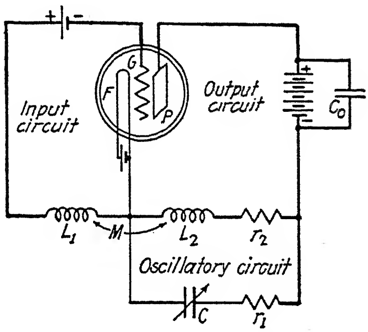


FIG. 148.—Oscillation.

inductance more than the critical values, the original disturbance will be increased in magnitude, its final amplitude being determined by these factors together with other constants of the circuit. The disturbance takes the form of an alternating current whose frequency is determined, except for minor corrections such as tube and circuit capacity, by the values of L_2 and C, as shown in Chap. VIII. The wave will not be sinusoidal unless precautions are taken to confine the working of the tube to the straight part of the characteristic.

The conditions necessary for oscillation are:

1. The tube must be capable of amplifying; that is, the power fluctuation produced in the output circuit must exceed that in the input circuit, so that a part of the output power may be fed back to the input, leaving some for the overcoming of losses.
2. The input current must be in phase with the output current.
3. A resonant circuit must be attached to the tube.

With these three requisites the tube constants together with the other constants of the circuit will determine the magnitude of the oscillations.

The frequency at which the tube will oscillate is, of course, determined by the resonant frequency of the oscillatory circuit, which, as shown in Chap. VIII, will be

$$f = \frac{1}{2\pi\sqrt{L_2C}}.$$

The range of frequency obtainable with vacuum-tube oscillators is almost unlimited. By employing a very large inductance and capacity and very close coupling by means of an iron core transformer, frequencies of but a fraction of a cycle per second may be obtained. On the other hand, other types of circuits have been made to oscillate with frequencies certainly as high as 600,000,000 cycles per second.

As in the case of amplifiers, it is usually desirable in vacuum-tube oscillators to work on the straight portion of the grid potential-plate current characteristic curve. Curvature in the characteristic introduces harmonics in the output current, which result in a waste of power.

Modulation.—In carrier-current telephony and telegraphy and also in radio transmission, the practice is to impress upon the line, or to radiate into space, a train of waves of such high frequency as to be inaudible. This is called the "carrier wave." According to the uses to which it is to be put, the frequency of the

carrier wave may be from about ten thousand up to some millions of cycles per second. It will be remembered from the chapters on sound that the range of frequency audible to the human ear may, for practical purposes, be considered as lying well below ten thousand cycles per second.

The most common source of such carrier waves is the vacuum-tube oscillator, the general features of which have just been discussed. Having solved the problem of sending out such waves as would constitute a proper carrier medium, two additional problems are involved: first, the modulation at the sending station of the carrier wave in accordance with the telegraphic signal or the sound waves which it is desired to transmit; and, second, the demodulation or the detection of the signal or sound variations at the receiving end.

Modulation consists in varying the amplitude of the otherwise steady high-frequency wave in accordance with the form of the low-frequency wave which it is desired to transmit. Thus, in speech transmsiison the successive changes in amplitude of the carrier wave are caused to follow, in their overall outline, the wave forms of speech. The frequency of the carrier wave is so high as to be well above the limit of audition; that of the speech wave is, of course, much lower and wholly within the audible range. The high-frequency wave carries the energy, the low-frequency wave modulates it and carries the intelligence.

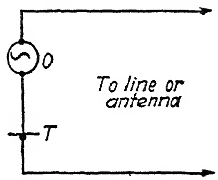


FIG. 149.—Simple case of modulation.

To show the modulation principle, without regard to vacuum tubes, we may take the very simple case shown in Fig. 149. In this, O is any generator of sinusoidal high-frequency current—the carrier-current generator. In practice it is usually a vacuum-tube oscillator such as has just been discussed. T is a telephone transmitter capable of varying the resistance of the circuit in the ordinary way. When the transmitter is quiescent the wave form of the current flowing is such as is represented at a (Fig. 150). Suppose that the diaphragm of the transmitter, after a period of rest, is given a sinusoidal vibratory motion, as by the sound waves of a tuning fork impinging against it. The movements of the diaphragm might then be represented by such a graph as b of Fig. 150, which shows a period of rest followed by two cycles of vibration again followed by a period of rest. This also may be

taken as representing the fluctuations in conductivity imposed on the circuit by the transmitter. Evidently, when the transmitter and generator are both operating, the transmitter, no matter what its frequency of vibration, will have no effect on the frequency of the carrier wave, but it will alter the amplitude of the carrier wave by imposing greater or lesser resistance in its path. The resulting wave form of the combined high- and low-frequency waves will be something like that shown in curve *c*. The carrier wave has the same frequency as before, but its amplitude has been made to vary in accordance with the sound wave.

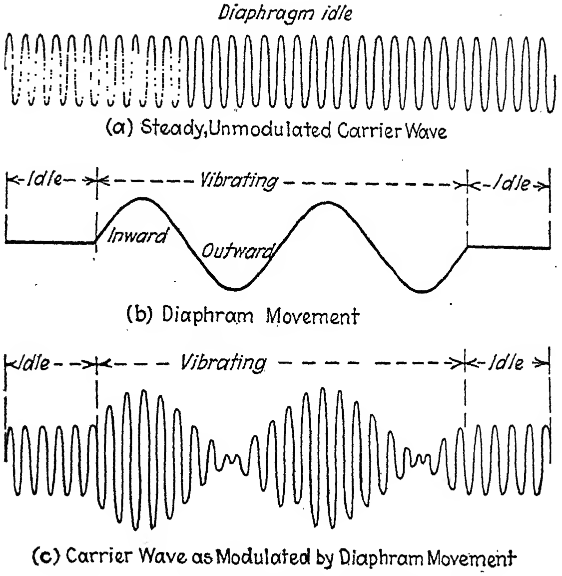


FIG. 150.—Modulation of carrier wave.

The arrangement of Fig. 149 would be rather ineffective because of the very small amount of power that an ordinary transmitter can control. It illustrates, however, in simple manner, how a high-frequency inaudible carrier wave may be modulated by varying its amplitude in accordance with the form of the audible voice or signal wave.

In order to attain more powerful modulation than can be had by causing the microphone directly to vary the resistance of the line or of the antenna circuit, as in Fig. 149, the vacuum tube is again called into play. In combination with the telephone

transmitter it forms by far the most effective means of modulation known. A single example of vacuum-tube modulation will suffice here. In the diagram of Fig. 151 the carrier wave from the oscillator and the voice wave from the telephone transmitter and amplifier are each impressed directly on the input circuit of the modulator tube. The wave form of the fluctuations of the grid potential is, therefore, that of a high-frequency carrier wave of constant amplitude superimposed on a low-frequency voice wave. We may think of the input wave as a low-frequency wave with carrier ripples of constant amplitude impressed upon it.

If the modulation tube were so arranged as to work on the straight part of its characteristic, the output wave might be amplified but would be of the same form, namely, a low-frequency wave with a high-frequency ripple. But what is desired is a high-frequency wave of varying amplitude. One way to obtain this is so to bias the grid of the modulator tube that the maximum negative value of the grid voltage reduces the plate

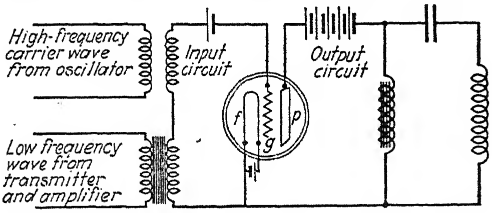


Fig. 151.—Principle of modulation by vacuum tube.

current nearly to zero. Then the ripples on the negative half of the wave will be amplified very little compared to those on the positive half, because of the comparatively small slope of the characteristic near its lower end. Consequently, the output will consist of a low-frequency wave on which is superimposed a high-frequency ripple which varies in amplitude in accordance with the low-frequency wave. A filter may then be inserted in the output circuit to suppress the low frequency, passing on the varying high frequency for amplification or transmission.

Such a filter could be simple, as indicated in Fig. 151. An iron core inductance of proper value would pass most of the voice-frequency current and hold back most of the high frequency. The condenser would hold back most of the low-frequency current and pass on the high to the coil which might be inductively coupled to the line or to amplifiers.

It can be readily shown either experimentally or by mathematical analysis that a modulated carrier wave may be resolved into a number of component frequencies, each of which is either the sum of, or the difference between, the carrier frequency and one of the component frequencies of the modulating wave. To illustrate: Take the very simple case of modulation shown in Fig. 150, where a sinusoidal carrier wave was modulated by a single sinusoidal modulating wave. Suppose the frequency f of the carrier wave to be 20,000 and that of the modulating wave f' 1,000 cycles per second. Then, ignoring such harmonics as might be introduced by irregularities, there would be three frequencies present in the modulated wave:

The carrier frequency, $f = 20{,}000$ cycles.
The summation frequency, $f + f' = 21{,}000$ cycles.
The difference frequency, $f - f' = 19{,}000$ cycles.

Similarly, if the carrier frequency were 1,000,000 cycles and the modulating frequency 1,000, as before, then the three frequencies in the modulated wave would be

$$f = 1{,}000{,}000.$$
$$f + f' = 1{,}001{,}000.$$
$$f - f' = 999{,}000.$$

It will be noticed in both cases that the original modulating frequency, $f' = 1{,}000$, has entirely disappeared.

In the case of speech communication the modulating wave, instead of being a simple sinusoidal wave of one frequency only, would be one of great complexity, consisting, as we have seen in Chap. V, of a number of simple sinusoidal components of various frequencies. The modulated wave in this case would contain, in addition to the carrier frequency, two bands of frequencies which may be referred to respectively as the "summation" and "difference frequencies," or as the *upper* and *lower sidebands*. If we are transmitting a comparatively narrow range of speech frequencies, say from 100 to 2,500 cycles per second on a carrier frequency of 20,000, then the frequencies of the upper sideband would be from 20,100 to 22,500 cycles and the lower sideband from 17,250 to 19,900 cycles. Again it is to be noted that the original voice frequencies 100 to 2,500 have entirely disappeared. The voice modulations are all represented, however, in each of the two sidebands. Either will suffice, and, as a result, one is often eliminated by means of a band filter from

the line or radio wave. Also, as will be shown in other chapters, the carrier frequency itself is sometimes eliminated from the wave that is actually transmitted, since of itself it carries no intelligence and can readily be replaced at the receiving station.

Demodulation or Detection.—When a modulated high-frequency wave arrives at a receiving station, whether wire or wireless, it has no power to actuate a telephone receiver, because all of its components are of such high frequency as to lie above the possible vibratory range of the receiver diaphragm. Even if the diaphragm could respond no good result would be accomplished, because the ear could not perceive vibrations of such high frequency. Some means must, therefore, be resorted to which will translate the intelligence-bearing components of the modulated wave from their high-frequency register down to such a low-frequency register as will enable the receiver to respond to them and the ear to interpret them. Here again the three-electrode vacuum tube has proved to be the best method of accomplishing an all-important result. It knows practically no limits of frequency.

The reason the telephone receiver diaphragm cannot respond directly to the modulated high-frequency wave is that the alternate pushes and pulls exerted by the wave occur so rapidly as to neutralize each other so far as any movement of the diaphragm is concerned. Before the diaphragm, with its relatively great mass, has time to move in response to an impulse in one direction, another impulse in the opposite direction neutralizes the first. It is obvious, however, that if the alternating wave could be rectified, that is, if the impulses could all be made to act on the diaphragm in the same instead of alternate directions, a deflection of the diaphragm would result. There would then be a succession of pulls following each other so rapidly as to constitute in effect one continuous pull. As this pull varied the diaphragm would move backward and forward in response to the variations.

This is essentially what is done in radio and in carrier-current wire telephony in order to make the inaudible modulated high-frequency wave audible. We have seen, in connection with Fig. 146, how, by working on the curved portion of the tube characteristic, a lopsided or partially rectified output current could be secured from a sinusoidal alternating input. The lopsidedness, or the degree of rectification, can be made greater

or less, according to the portion of the tube characteristic worked on. It is evident that the upper output wave X' (Fig. 146) would cause no appreciable deflection of the diaphragm while the lower one Y' would. If the amplitudes of the lopsided high-frequency wave Y' were made to vary, the variations occurring slowly enough to be within the audible limit, it is

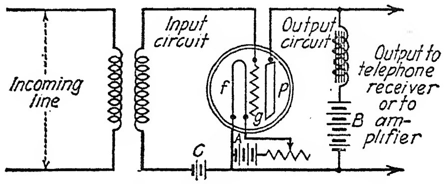


FIG. 152.—Demodulation circuit.

evident that the diaphragm would move back and forth in response to the variations and would emit corresponding sounds.

Figure 152 shows a demodulator circuit, working on this principle, as it is used in multiplex carrier-current telephony over long-distance wire lines. In this case the output is delivered to an amplifier for further amplification. The same circuit would work in radio reception, in which case the primary of

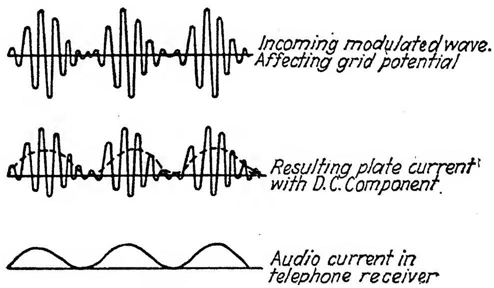


FIG. 153.—Modulated and demodulated waves.

the input coil would be connected in the antenna circuit instead of in the line.

Figure 153 shows roughly the wave characteristics of the input and output, and of the direct-current component of the latter, as it would be felt by a telephone receiver if connected directly across the output circuit of Fig. 152.

In this chapter only one or a few specific circuits and methods of operation have been dealt with in connection with each of the

more important applications of the vacuum tube, the purpose here being to deal mainly with the principles of operation and the characteristics of the tube itself and to show its wide field of application in electrical communication.

Four- and Five-electrode Tubes.—In order to overcome the electrostatic effects of plate potential changes on grid potential, which are especially troublesome at high frequencies, a second grid is sometimes inserted between the control grid and plate. This "screen grid" is held at a fixed potential, usually somewhat less than the average plate voltage, and effectively screens the control grid from the effects of variations in plate potentials. The screen grid also serves to control the space charge in such a way that under best conditions a tube of this type may have a voltage amplification of several hundred. This type of tube has very high plate resistance and hence is not generally effective as a power amplifier. By the insertion of still another grid it seems to be possible to retain the high voltage amplification and at the same time reduce the plate impedance so that the tube will handle a considerable amount of power. The five-electrode tube thus constituted is, however, in its early stages of development, and its full possibilities have probably not been realized.

CHAPTER X

MAGNETIC MATERIALS

Of all the elements only three show magnetic properties in sufficient degree to be ordinarily classed as magnetic materials. These are the metals, iron, nickel and cobalt. Of the three, iron, either by itself or in its various forms of steel, far surpasses the others in magnetic properties. In fact, until a very short time ago, nickel and cobalt could be practically dismissed from the list of materials that were of any importance whatever for magnetic use, leaving iron alone in the field. It is fortunate that the metal which far outranks all others in magnetic importance should also be the cheapest, and, probably,—all things considered—the most easily worked.

Pure iron has long been known as the material most easily magnetized. Even such an authority as J. A. Ewing, author of "Magnetic Induction in Iron and Other Metals," an outstanding work on ferromagnetism, had concluded that no material would ever be found which would surpass the best grade of iron in this respect.[1] Notwithstanding this general belief, Mr. G. W. Elmen, of the Bell Telephone Laboratories, has, succeeded in producing a series of alloys which are very much more easily magnetized, and which, in other respects particularly pertinent to telephony, are far superior to the best grades of iron or steel. As nickel or cobalt, or both, are, with iron, the constituent elements of these alloys, it is evident that we may no longer dismiss these humbler members of the magnetic group as of no importance for magnetic purposes.

A new aspect is thus given to the old subject of magnetic materials, and the important place which iron has long occupied as one of the fundamental materials of telephony bids fair to become even more important by its closer association with the long-neglected members of its family, nickel and cobalt. The placing in the hands of the telephone engineer of new magnetic

[1] McKeehan, L. W., A Physical Background for Permalloy, *Bell Laboratories Record*, December, 1926.

materials that are, under the conditions peculiar to his art, far superior to anything heretofore available must be of profound significance. The results to be expected are only at their beginning, and it is impossible to state how far or into what fields they will extend. It seems almost a foregone conclusion that these new magnetic materials, and the ones which further research will surely bring forth, will exert an ever increasing influence in the design of such apparatus as telephone receivers, induction coils, loading coils, and relays, and will work revolutionary changes in the construction of some of the wires which carry telephone currents.

In order to understand the principal magnetic characteristics possessed in varying degree by different magnetic materials, a little elementary discussion is desirable.

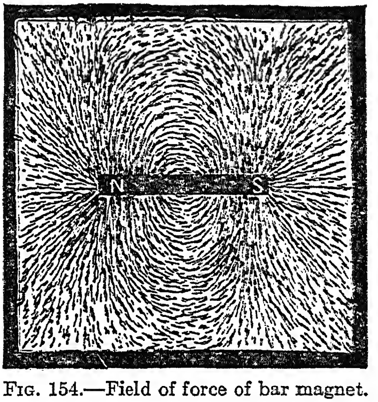


Fig. 154.—Field of force of bar magnet.

The phenomenon which we know as "magnetic field" is a property which is distributed continuously through space. We will find it convenient, however, to consider the effect as grouped into lines of force, each representing a unit of field strength. The number of such lines passing through a unit area will then be a measure of the strength of the field.

Magnetomotive force or magnetizing force is the force which tends to set up a magnetic field. Since we conceive a magnetic field to consist of magnetic lines of force, we may say that magnetomotive force is the force which tends to set up magnetic lines of force. In the case of a bar magnet, for instance, the lines of force are conceived as passing through the length of the magnet, emerging from one of its poles and passing in closed curves through the external circuit of air, or other material, back to the other pole where they reenter the bar. A magnet may be made to map out its own field of force by placing it in a horizontal position and dropping iron filings on a sheet of paper laid just above it. Figure 154 shows such a map, produced by the bar magnet *NS*.

The total number of magnetic lines set up by a magnetizing force is called the "magnetic flux." The path followed by the flux is called the "magnetic circuit." Thus, in Fig. 154, the total number of lines passing through the bar, or the total number of lines passing through the air in returning from one pole to the other, constitute the magnetic flux; and the path followed by these lines through the bar and back through the air constitutes the magnetic circuit.

Magnetomotive force is related to magnetic flux as cause and effect. Thus a given magnetomotive force will cause a certain magnetic flux in a magnetic circuit just as a given electromotive force will cause a certain current flow in a given electric circuit.

The amount of magnetic flux will be proportional to the total magnetizing force and inversely proportional to the *reluctance* of the magnetic circuit, reluctance being the property of the magnetic circuit analogous to resistance in an electric circuit. We have, then, a sort of Ohm's law for magnetic circuits; for, when expressed in proper units, the magnetic flux is equal to the total magnetizing force divided by the reluctance, or,

$$\text{Magnetic flux} = \frac{\text{magnetomotive force}}{\text{reluctance}}.$$

The reluctance of a magnetic circuit is dependent on the geometry of the circuit, that is, on its physical dimensions and form, and also on the character of the material of which the circuit is made. At the present moment we may dispose of the geometry of the circuit by saying that, like the resistance of an electric circuit, the reluctance of a magnetic circuit through a given material is proportional to the length of the circuit and inversely proportional to its cross-sectional area. Aside from physical dimensions of the magnetic circuit, a characteristic of the material of which the circuit is formed determines the reluctance. This is one of the characteristics of the material itself with which we are particularly concerned in this chapter.

To illustrate: A coil of wire with an air core (a solenoid) will exhibit magnetic properties upon the passage of an electric current through it. The magnetizing force due to the current will set up a magnetic flux threading the coil. If a bar of iron or a bundle of iron wires be placed in the coil, the magnetic effects produced by the same current will be enormously increased. This is because the same magnetizing force applied to iron will

produce a far greater flux than if applied to air or other non-magnetic material.

This peculiar characteristic of a material which permits a given magnetizing force to set up a greater or less magnetic flux through it is called "permeability." It is analogous to the conductivity of a material, which permits a greater or less current of electricity to flow through it in response to a given electromotive force. The permeability of a substance is evidently a measure of the ease with which it may be magnetized. More specifically, it is the ratio between the total flux through the material and the magnetizing force which causes the flux.

In order to facilitate specific comparison of permeabilities and also of other qualities of magnetic materials, it is customary to consider, instead of the total magnetizing force and the total flux, only that part of each which is applied to a unit area of the cross-section of the material. When made thus specific to a unit area, the magnetizing force applied to a square centimeter of the cross-section of a core, for instance, is called the *intensity of magnetizing force;* and the resulting flux in the same area is called the *flux density.* Obviously, in a core of one square centimeter cross-section, the total flux would be numerically equal to the flux density and the total magnetizing force to the intensity of magnetizing force.

It is customary to refer to the intensity of magnetizing force by the symbol H; to the flux density produced by the magnetizing force by the symbol B; and to the permeability of the material by the Greek letter μ. If we consider a core with a cross-section of one square centimeter we have, then, from the definition for permeability just given,

$$\mu = \frac{B}{H}.$$

Obviously, the ratio will be the same for a core of any other cross-section in which the distribution is uniform, for both B and H will be multiplied by the same area in order to determine the total flux and the total magnetizing force.

The terms "magnetic flux" and "magnetic flux density" must be carefully distinguished. Magnetic flux, usually indicated by the symbol ϕ, means the total flow of lines of force through a magnetic circuit. The magnetic flux density, B is the number of magnetic lines per unit area of the cross-section

of the iron or other material. Evidently, then, $\phi = Ba$, where a is the area.

A few words regarding magnetic units are necessary. The unit of intensity of magnetizing force or of strength of field—that is, the unit in which H is measured—is the "gauss." It is a field of such strength that a unit pole situated in it experiences a force of one dyne. Expressed in another way, an intensity of magnetizing force of one gauss is such an intensity as would set up one line of force per square centimeter in air. Thus we may express the intensity of a magnetizing force H either in terms of the number of lines of force per square centimeter it would produce in air or merely in terms of so many gauss.

No single word unit seems to have been authoritatively adopted by which to measure B, the magnetic flux density. There is a unit of magnetic flux, the "maxwell," which is the equivalent of one line of force. This is comparatively little used in practice because it is shorter and more expressive to say a "flux of so many lines" than "a flux of so many maxwells." When it comes to expressing B, the magnetic flux density, it becomes necessary, in order to comply with strict scientific sanction, to do so in terms of the number of lines per square centimeter or, its equivalent, the number of maxwells per square centimeter.

While the gauss, under strict interpretation of scientific definition, is applicable only to intensity of magnetizing force H, it has become common practice to use it also as applying to lines of induction per square centimeter, that is, to magnetic flux density B. It is not uncommon to find authorities defining the gauss as referring to H and immediately afterward applying it to B. In much of the modern writing on the subject, gauss is used perhaps more frequently as measuring B than as measuring H. This departure of authoritative practice from authoritative definition is likely to confuse the mind of the reader unless he is forewarned; when understood it leads to no confusion. In the following text the endeavor will be made to use the word "gauss" in reference to H and the longer term "lines per square centimeter" in reference to B. In some of the figures, however, taken from authoritative sources, it will be noticed that gauss is interchangeably used as applying sometimes to B and sometimes to H.

Since we measure the intensity H of a magnetizing force by the number of lines of force it will set up across an area of one

square centimeter of a circuit composed entirely of air, it is evident that for circuits composed of air, B will be numerically equal to H. The permeability μ of air is therefore unity. This leads to another definition of the permeability of a substance: The ratio of the flux that a given magnetizing force will set up in that substance to the flux that it will set up in air.

The permeability of air, or of any other non-magnetic substance, is constant and, as has just been said, is taken as unity. The permeability of a magnetic substance, however, is not constant; it varies with the flux density and also with the previous history of magnetization. For these reasons graphs—called

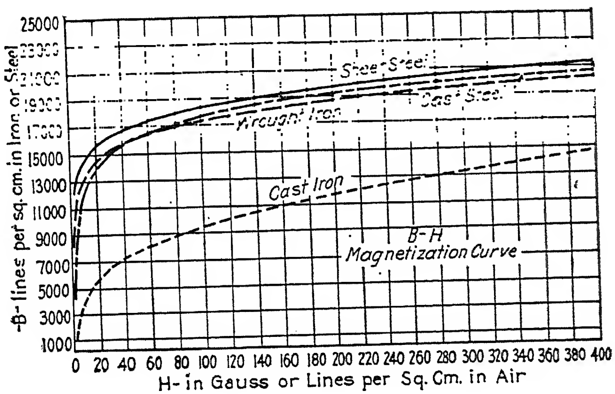


FIG. 155.—Magnetization or B-H curves.

"magnetization curves"—are employed to show the amount of flux which a given magnetizing force will produce in the kind of material to which the curve relates. These can be determined only by experiment, but, once determined, they afford a convenient means of predicting the flux density that any magnetizing force will produce.

In Fig. 155 are shown characteristic curves of the relationships between the magnetizing force and the resulting magnetism for cast iron, wrought iron, cast steel and sheet steel. In these the abscissæ show the intensity of the magnetizing force H expressed in terms of the number of gauss, or of lines per square centimeter which it would set up in air. The ordinates represent the corresponding flux densities B; that is, the actual induction in lines per square centimeter for each value of the magnetizing force. Such curves as this are commonly called "magnetization

curves" or, frequently, merely "*B-H* curves," since they show the relationship between *B* and *H*.

These particular curves of Fig. 155 are of value principally as showing in a very general way the characteristics of the different classes of iron and steel indicated. As will be shown, slight variations in composition, due perhaps to impurities, and different degrees of hardness, due to processes of manufacture or subsequent heat treatment, as well as other factors, will produce wide variation in magnetic characteristics, so that for

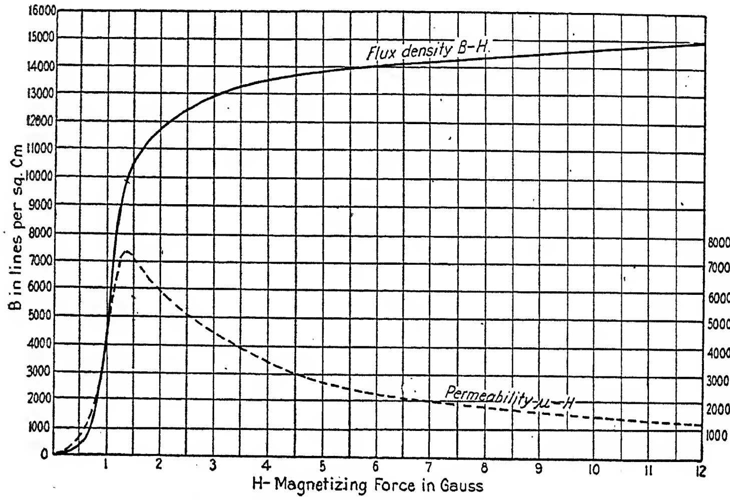


FIG. 156.—Magnetization and corresponding permeability curves.

actual quantitative use the only safe way is to use curves which have been determined for the particular grade of material in question.

Moreover, for the purposes of the telephone engineer, who often works with very small magnetizing forces and very low flux densities, curves drawn on the scale employed in Fig. 155 would be of little value even if applicable to the exact magnetic material in question. This will be apparent when we inquire more particularly into the form of such curves at their lower ends, corresponding to low values of magnetizing force and flux.

When the horizontal scale is magnified, so as to make more apparent the effect of small changes in magnetizing force, the curve is seen to take the general form of the solid line of Fig. 156.

This we may consider in three sections: For very small magnetizing forces the flux is small and increases slowly as the magnetizing force increases. Here the ratio of B to H is comparatively small, indicating relatively low permeability. Then, in the section just beyond, the curve rises rapidly. Small increments of magnetizing force produce large increments in flux. The ratio of B to H increases rapidly to a maximum along this portion. Beyond this the curve begins to flatten out, indicating the approach to saturation. From this point on, the ratio of B to H will become less and less, indicating gradually decreasing permeability.

The permeability for different degrees of magnetization may be readily calculated from the magnetization curves. Thus, for example, the sample of wrought iron characterized in one of the curves of Fig. 155 requires a magnetizing force of 30 lines per square centimeter to produce a flux density of 15,000 lines per square centimeter. The permeability μ at that point on the curve is, therefore, 500. For a value of H of 10 the value of B is about 13,000, indicating a permeability at that point of 1,300. Similarly, the permeability may be calculated for any other point on the magnetization curve, and it will be noted that beyond the knee of the curve the permeability rapidly decreases. For instance, at a magnetizing force of 400 gauss the flux density is 20,000, indicating that the permeability has fallen to 50.

For convenience we may plot a curve showing the value of the permeability of a piece of magnetic material for each different value of the magnetizing force, or for each different value of the flux density, by using the values of either H or B on the horizontal scale and the corresponding values of μ on the vertical. The values of μ have thus been plotted against the corresponding values of H for the B-H curve of Fig. 156, resulting in the dashed line curve on that figure. Any ordinate on this curve represents the ratio B/H for the corresponding point on the B-H curve.

The final flattening out at the higher densities, noticeable in all magnetization curves (Figs. 155 and 156), and the corresponding lowering of permeability, shown by the permeability curve of the latter figure, indicate an important tendency to "saturation" at high densities, possessed in greater or lesser degree by all magnetic materials. Actual saturation never occurs, since the point is never reached where an increase in magnetizing force will not cause some increase in the flux density. Practically,

however, saturation does occur, since a point is reached beyond which the small increase in flux to be gained by further increase in magnetizing force is not worth while. In common practice, therefore, even where high densities are required, it is usually not economical to work far beyond the knee of the magnetization curve.

In addition to permeability, there are other characteristics of magnetic materials which are of varying degrees of importance according to the use to which the material is to be put. Among these are *retentivity*. This is the ability of a material to retain residual magnetism, that is, to remain in a magnetized condition when the magnetizing force has been removed.

Two manifestations of retentivity are familiar to everyone at all familiar with telephone apparatus. One is the tendency of a telephone or telegraph relay to stick, or, in other words, to continue to hold its armature after the current through its coil has ceased. This is an undesirable manifestation and is to be avoided, as far as possible, by the use of a grade of iron which has the quality of retentivity in as low degree as possible. The other common manifestation of retentivity is that of the permanent magnet. This, after removal of the force which originally magnetized it, will hold its magnetism in pronounced degree for many years. Obviously, in the production of permanent magnets the quality of retentivity in high degree is to be sought.

The retentivity of a magnetic material is measured by the residual flux which remains in a unit of area of its cross-section after the magnetizing force has been removed. Evidently the retentivity of air or other non-magnetic materials is zero, for no magnetic flux remains in them when the magnetizing force is withdrawn.

Closely allied to retentivity is *coercive force*. It is a measure of the power to resist demagnetization, and it is defined as the value of the magnetic field necessary to reduce the intensity of magnetization to zero. As related to retentivity, therefore, it is the magnetizing force required to destroy a residual field of force due to retentivity. In air, of course, it does not exist because when the magnetizing force is withdrawn the residual field automatically disappears. In iron, however, it is always found that in order to destroy the residual field which remains after the withdrawal of the magnetizing force, some magnetizing force in the opposite direction must be applied. Coercive force, like permeability, is a

variable for a given piece of magnetic material and depends on the previous history of magnetization.

Retentivity and coercive force are thus distinct though closely related attributes of magnetic material. Retentivity relates to the density of flux that will be retained by the material after the withdrawal of the magnetizing force, while the coercive force is the magnetizing force required to take away this retained flux. Thus it may be possible for each of two kinds of magnetic material to retain the same amount of magnetism. One of these may have a feeble hold on it, so that a small coercive force will remove it, while the other, having a stronger hold, will require a larger coercive force to remove it.

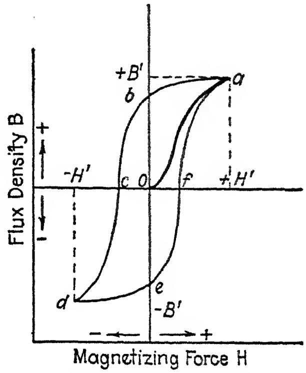


FIG. 157.—Typical hysteresis curve.

Still another characteristic of magnetic material is *magnetic hysteresis.* This is a sort of magnetic friction within the molecular structure of the material, which causes the magnetizing effect to lag behind any changes in magnetizing force. It is closely related both to retentivity and to coercive force, as will be shown.

We may gain an idea of the relationship between retentivity, coercive force and hysteresis by a consideration of Fig. 157. In this, as in the magnetization curves already considered, ordinates represent flux density and abscissæ magnetizing force. Here, however, the magnetizing force is carried through a cycle from zero to a positive value $+H'$, then through zero again to an equal negative value $-H'$ and back again to zero. This is such a cycle of magnetizing force as would be exerted on the core of an electromagnet while energized by an alternating current. As a result of this alternating magnetizing force, the resulting flux alternates between values of $+B'$ and $-B'$.

As a preliminary to the consideration of this cycle we will start with the iron in a completely unmagnetized condition. Then, as we plot the values of magnetizing force and resulting flux density as the magnetizing force is increased from zero to $+H'$, the heavy curve Oa will be formed. This is evidently an ordinary B-H curve of the kind already considered. If we slowly decrease

the magnetizing force from $+H'$ to zero, noting the corresponding values of B on the return journey, a different sort of curve will result. Instead of following the same path aO during demagnetization, a new path ab will be traced. This means that when the magnetizing force has been reduced to zero, the flux density will not have decreased to zero, but that a residual flux having a value represented by the ordinate Ob will remain. Evidently the retentivity of the iron has resulted in this flux Ob remaining after the magnetizing force has been withdrawn.

To destroy this residual flux will require a negative magnetizing force, and it will not be until the magnetizing force has been reversed and reached a value Oc in that direction that the flux will reach its zero value. Oc is then the coercive force necessary to destroy the residual flux Ob caused to remain by the retentivity of the iron.

Following the cycle still further, the section of the curve cd is formed as a result of increasing the negative magnetizing force from c to its maximum negative value $-H'$, the maximum negative flux density $-B'$ being reached at that time.

During the other half cycle, while the magnetizing force is being reduced from its maximum negative value $-H'$ to zero and again increased to its maximum positive value $+H'$, the process will be just the reverse of that already traced. Reduction of the magnetizing force from $-H'$ to O will result in the section de of the curve, the flux not falling to zero but remaining at a value $-Oe$, due to the retentivity of the iron. This residual negative flux will be destroyed along the section ef of the curve, the flux reaching its zero value only when the magnetizing force has reached a positive value Of. As the magnetizing force is still further increased to its maximum positive value H', the positive flux density is built up along the line fa.

Such a closed curve, representing the varying values of flux density as the magnetizing force alternates between positive and negative maximum values, is called an "hysteresis loop." Evidently the width of the loop, on its horizontal axis, is equal to twice the coercive force; and its height, on its vertical axis, is equal to twice the residual magnetism held by retentivity.

The fact that the curve of the loop (Fig. 157) does not follow the original B-H curve, either on its return from a, during demagnetization, or its reapproach to a, during remagnetization, is significant. The degree by which it fails to do so is measured

by the area enclosed within the loop. This area is a measure of the loss of energy which occurs, presumably due to the molecular rearrangement within the structure of the iron during the progress of one magnetic cycle. This loss, called "hysteresis loss," is due supposedly to molecular friction; and, since it is spent in heat, it is lost so far as any useful results are concerned. Hysteresis loss is one of the so-called "iron losses" always present in some degree where alternating or variable currents cause magnetic variations in iron, steel or other magnetic materials.

Another characteristic of magnetic material is its *resistivity*. This is the reciprocal of its electrical conductivity and is the same as specific resistance. While it is a purely electrical characteristic, it has an important bearing on the suitability of a material for certain magnetic purposes.

Rapid changes of current in a coil tend to induce other currents in all near-by conductors. Necessarily, the iron in the cores and other parts of the magnetic circuits of electromagnetic apparatus is closely associated with the current-carrying coils, and, in so far as this iron is a conductor, it offers a path for such induced currents. Currents thus induced in the magnetic material of electromagnetic apparatus are called "eddy currents." Their energy is spent in a useless heating of the material. Eddy-current losses form the second of the so-called "iron losses"; hysteresis, as just stated, being the other. Obviously, the higher the specific resistance or resistivity of the iron the smaller will be the eddy current losses.

From the foregoing the characteristics which govern the choice of magnetic material for different uses are seen to be permeability, retentivity, coercive force, hysteresis and resistivity. These qualities assume varying degrees of importance according to the use to which the material is to be put.[1] From the standpoint of adaptability we may very roughly divide magnetic materials into two classes:

1. The so-called "magnetically soft" materials. These are easily magnetized and demagnetized. In these the qualities generally desired are high permeability, low retentivity, low coercive force and low hysteresis loss. Sometimes high resistivity

[1] For some of the following information in this chapter regarding the composition and characteristics of the different grades of magnetic iron and steel, I have drawn freely on a short article, Magnetic Materials, by I. C. Pettit in the *Bell Laboratories Record*, January, 1927.

is also an important consideration. The soft materials are used, for instance, as the cores and other parts of the magnetic circuits of electromagnets and transformers, and, in general, wherever it is desired to have the magnetic flux rise and fall as nearly as possible in accordance with the current which develops it.

2. The so-called "magnetically hard" materials. These are hard to demagnetize. The desiderata, therefore, are high retentivity and high coercive force. They are used in permanent magnets such, for instance, as those which furnish permanent magnetic fields to telephone receivers, magneto generators and polarized relays.

Generally speaking, the magnetically soft materials include wrought iron, soft low-carbon steel, electrolytic iron and the various recently developed alloys of iron with nickel or cobalt or both. "Swedish iron," and "Norway iron" were once largely used for magnet cores and similar purposes, but these grades have now been largely superseded by grades of soft wrought iron known as "magnetic iron," "Armco" iron (which is a very pure soft iron) and silicon steel, a steel containing about 4 per cent of silicon.

The magnetically hard materials are all of hardened steel. A relatively small carbon content gives high coercive force, and the addition of chromium, tungsten, molybdenum or cobalt in certain proportions serves to increase the magnetic hardness, particularly the coercive force. Certain grades of manganese steel are almost non-magnetic, but, as a rule, very small proportions of manganese are not seriously harmful. Among the grades of steel used for permanent magnets are carbon-manganese steel, chrome steel, tungsten steel and, recently, high-cobalt steel.

The choice of magnetic material from either group is governed by the particular characteristics desired for the proposed use, cost often being an important consideration. A few examples may illustrate this.

"Magnetic iron" is a suitable grade to use for the cores and armatures of most electromagnets which operate on direct current. Such magnets are made in very large quantities so that the cheapness of this material is important. Its permeability even at fairly high flux densities is rather high, which is necessary in order that the magnetizing current in the coil may

create an effective pull. Its retentivity and coercive force are fairly low which means that there will not be an undue amount of "sticking" due to residual magnetism when the energizing current ceases to flow. Its resistivity is low, but this is not an objection for ordinary direct-current use, since eddy-current losses are then of small moment. For these reasons, and particularly on account of its cheapness, magnetic iron is largely used for direct-current magnet cores.

There are many exceptions to this practice, however, as, for example, in the case of a relay magnet where especially quick action in the attraction and the release of the armature is demanded. In such a case, while the energy loss, as such, may be unimportant, the establishment of an eddy current each time the circuit is made or broken causes a time lag in the response of the armature. This is not, in the case of release, the ordinary sticking of the armature due to retentivity but rather to the actual magnetic effect of the eddy current itself, which tends to prevent a change in the pull of the core.

Other exceptions to the use of magnetic iron for direct-current magnet cores are in cases where extremely high magnetic efficiencies are required. Here, by using some of the high-permeability alloys, larger flux densities may be created by the limited available magnetizing force, resulting in stronger magnetic pulls than could otherwise be attained.

Where alternating currents, particularly voice currents, are involved, silicon steel finds wide use. It also is comparatively cheap and its resistivity is about five times that of magnetic iron. This greatly reduces the eddy-current losses, which are further reduced by laminating or otherwise subdividing the cores to break up the eddy-current paths. Silicon steel also has higher permeability at small flux densities than magnetic iron and is, therefore, more effective in its response to small changes in weak magnetizing forces.

As examples of the newer magnetically soft materials two of the "permalloys" may be mentioned at this point by way of comparison with the older materials. The "45 permalloy" (45 per cent nickel 55 per cent iron) has a permeability at low and moderately high flux densities about five or six times as great as silicon steel, with a resistivity almost as high and a coercive force considerably higher. The "78.5 permalloy" (78.5 per cent nickel 21.5 per cent iron) has about twenty times

as high a permeability at low and moderately high flux densities as silicon steel, with a resistivity almost as high and a coercive force only one-tenth as great. Figure 158 shows comparative permeability curves for this and silicon steel for various values of flux density. Further consideration of these and other new magnetic materials will be given later in this chapter.

In all of the soft magnetic materials much depends on the heat treatment that is given after the final operations employed in fabrication. The processes of rolling or drawing the sheets or

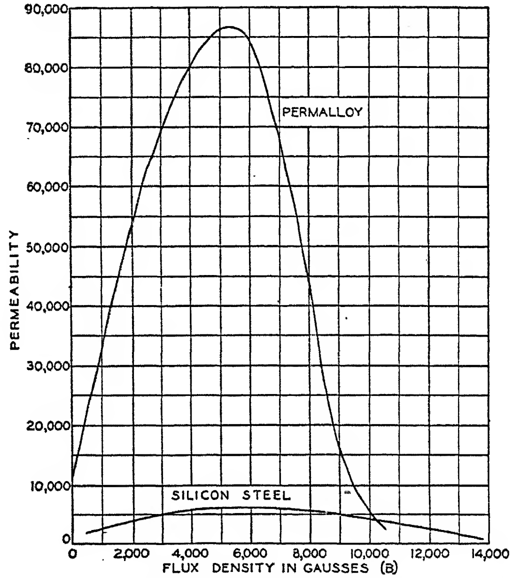


FIG. 158.—Permeability—flux-density curves for permalloy and silicon steel. (*Courtesy of Bell Telephone Laboratories.*)

rods, and the subsequent operations of cutting or otherwise shaping them into the desired forms strains and distorts the crystalline structure. The general tendency of this alteration is to lower permeability and raise coercive force. In the ordinary soft materials, like magnetic iron and silicon steel, the final process in manufacture is to heat the parts to a given temperature and for a given time, while sealed in metal boxes, and then allow them, while so sealed, to cool very slowly. This anneals them and secures the final magnetic characteristics. In the newer magnetic alloys heat treatment has even more marked effects on magnetic qualities than in the ordinary irons and

steels, and in some of them it acts in quite the opposite direction. In these alloys, therefore, the requirements for heat treatment are even more exacting, as will be more fully pointed out later in this chapter.

The magnetization and demagnetization curves for several of the permanent magnet steels, carbon-manganese steel, tungsten steel and cobalt steel are shown on a comparative basis in Fig. 159. The higher coercive force for the cobalt steel is to be noted.

Carbon-manganese steel and 1 per cent chrome steel are practically equivalent and are made by the cheapest steel-making processes. Although not furnishing as high permanent flux

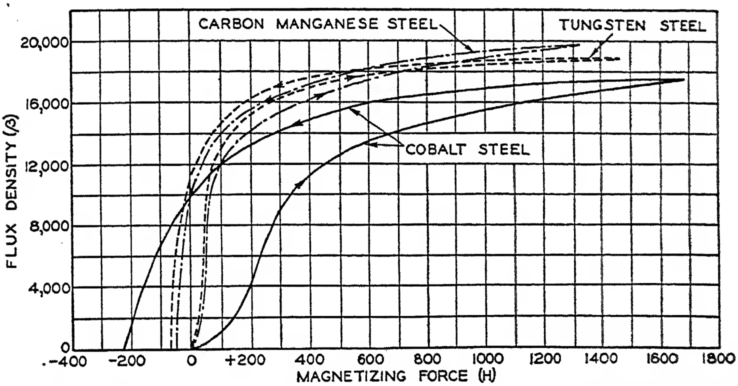


FIG. 159.—Magnetization and demagnetization curves for permanent-magnet steels. (*Courtesy of Bell Telephone Laboratories.*)

densities as other grades, their relative cheapness makes it practicable to supply the necessary strength of permanent field by using large magnets in cases where economy of space is not a controlling item. They are, therefore, largely used in making the permanent magnets in such telephone apparatus as magneto-generators and ringers.

Tungsten steel, since it contains about 5 or 6 per cent of tungsten and is necessarily made by the crucible process, is about four or five times as costly as carbon-manganese or 1 per cent chrome steel. As shown by the curve of Fig. 159, however, it has a considerably higher coercive force, and for this reason is used in such telephone apparatus as head receivers and polarized relays, and in other places where small bulk and short magnets are necessary.

High-cobalt steel is one of the recent important developments in magnetic materials. Its composition is approximately 35 per cent cobalt, 8 per cent tungsten and 3 per cent chromium, with carbon content about the same as in other steels. Its coercive force is remarkable, about three or four times that of tungsten steel. On account of its great power to retain its magnetism, it is destined to find increasing use in telephone apparatus. It is now used in some telephone receivers, in electromagnetic phonograph reproducers and in high-speed polarized relays. Its present cost is about ten times that of tungsten steel.

As stated, all of the permanent magnet steels are given a hard temper before being magnetized. Only in this condition do they acquire the necessary coercive force to assure permanency. Much depends on this temper, which is given by heating the magnets, after their fabrication, to a definite temperature which is above red heat and then suddenly quenching them in water or oil.

On account of its far-reaching importance, as well as its great scientific interest, some further account will be given of the results secured by Elmen in his investigation of the magnetic properties of the iron-nickel-cobalt alloys. The following information concerning these alloys has been taken largely from an article[1] just published by him, to which the reader is referred for much additional matter.

The iron, nickel and cobalt from which the alloys were made were the purest commercial grades obtainable, Armco iron, electrolytic nickel and a high-grade cobalt containing only slight impurities. While some of the impurities affected the magnetic properties unfavorably, no attempt at further refinement was made because the principal interest lay in determining alloys that could, if desired, be reproduced commercially on a large scale.

Since the magnetic properties of the various alloys were peculiarly dependent upon heat treatment, after much experimentation three standard heat treatments were adopted. Several samples of each alloy, after their fabrication into the desired standard ring shape for testing, were packed in a pot and heated

[1] Elmen, G. W., Magnetic Alloys of Iron, Nickel and Cobalt, *Journal of the Franklin Institute*, vol. 207, pp. 583–617, May, 1929; also in *The Bell System Technical Journal*, vol. 8, pp. 435–465, July, 1929.

to a temperature between 900 and 1000° C., and, after being held at this temperature for one hour, were allowed slowly to cool in the pot to room temperature. Samples receiving only this treatment were referred to as "annealed." Some of the annealed samples were again heated to 600° C., then removed and cooled comparatively rapidly on a copper plate in air. Samples receiving this additional treatment were referred to as "air quenched." Some of the annealed samples were heated to a lower temperature, 425° C., and held there for 24 hours. These samples were referred to as "baked." As will be shown, samples of exactly the same composition, cast from the same ingot, showed markedly different magnetic properties according to whether they had been annealed, air quenched, or baked.

For each sample, the flux density B was determined for magnetizing forces, H extending from a few thousandths of a gauss up to 1,500 gauss, and corresponding magnetization and permeability curves were plotted. Hysteresis measurements were also made on each sample, ordinary hysteresis loops being made for flux densities varying between plus and minus 5,000 lines per square centimeter. For alloys which seemed to warrant further investigation, hysteresis loops were determined for different maximum flux densities from less than 100 lines per square centimeter, up.

In graphically illustrating the magnetic properties of these alloys, Mr. Elmen uses effectively the "solid diagram" which has long been used in representing other physical properties, such as tensile strength of binary and ternary alloys. The base of such a solid diagram is the equilateral-triangle composition diagram of Fig. 160. In this the undiluted metals—iron, nickel and cobalt—are represented by the corners, as indicated. The binary alloys are represented by points along the respective sides, the numbers along the sides indicating the percentage of the metal written along the side. Thus the point marked *a* would indicate 78.5 per cent nickel and, since the point is on the nickel-iron side, 21.5 per cent iron. The ternary alloys are represented by points within the triangle, the three coordinates, in the direction of the arrows, pointing to the corresponding percentages of each metal. To illustrate: The point marked *b* indicates by its position in the triangle an alloy of 50 per cent nickel, 20 per cent cobalt and 30 per cent iron. The dots on Fig. 160 represent the alloys particularly covered by this investi-

gation and give an idea of the completeness with which the field was covered.

This explanation of the flat composition diagram of Fig. 160 is necessary to understand the solid diagrams now to be referred to. If we erect upon each point on the triangle an ordinate perpendicular to its plane and make each of these ordinates proportional in height to some particular property, such as the permeability of the corresponding alloy, then a surface drawn through the tops of all the ordinates will represent by its elevation at any point the value of the property of the alloy represented by that point.

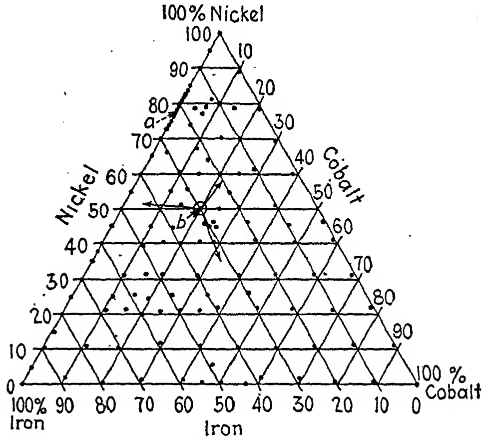


FIG. 160.—Composition diagram for nickel-iron-cobalt series.

In each of these solid diagrams we must confine our attention to but one particular property of the various alloys prepared under one particular set of conditions. A diagram of initial permeabilities for annealed alloys, for instance, will be quite different from one of maximum permeabilities for the same alloys. Likewise, diagrams of any particular magnetic quality are likely to differ widely for annealed, air-quenched or baked alloys. The variations are practically infinite, and only by confining ourselves to one set of conditions at a time may confusion be avoided.

Initial permeability is an indication of permeability at very low densities. It is determined by extending the permeability curve by projection to a point where it crosses the permea-

bility axis. The intercept on that axis represents the initial permeability.

Figure 161 shows a solid diagram representing this particular quality, initial permeability, for all annealed binary and ternary alloys. In this the left-hand corner represents 100 per cent iron, the right 100 per cent nickel and the front 100 per cent cobalt.

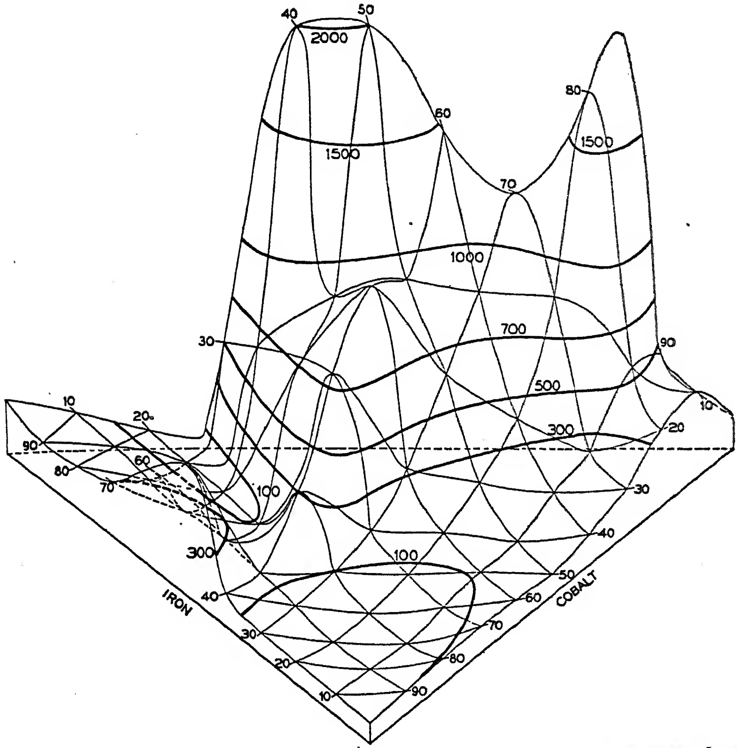


Fig. 161.—Initial permeabilities for annealed alloys. (*Courtesy of Bell Telephone Laboratories.*)

The heavy lines on the figure are contour lines, such as are used on topographical maps, but, in this case, the contours indicate lines of equal initial permeability. The numbers adjacent to these lines show the initial permeabilities represented by the lines.

It is seen that for annealed alloys the greatest initial permeabilities occur on the back edge of this diagram, which represents the nickel-iron alloys. Starting at the left-hand corner and following the curve of this back edge, it is seen that the initial

permeability of 100 per cent iron is something less than 300. For small additions of nickel, up to about 28 per cent, this value decreases to less than 100, but from that point there is a rapid rise as nickel is added until at about 45 per cent nickel the initial permeability reaches a maximum of over 2,000. Beyond 50 per cent nickel the permeability decreases to about 1,200 at 70 per cent and then increases again to a second maximum of about 1,900 at 83 per cent nickel. From 83 per cent on the permeability drops rapidly to its value of about 200 for 100 per cent nickel.

Other regions of interest on this diagram might be pointed out but only the hump on the iron-cobalt plane between 40 and 70 per cent iron will be mentioned. This rise indicates a maximum initial permeability of about 600 along this range of this binary alloy—more than twice that of Armco iron.

Plotting the initial permeabilities for air-quenched instead of annealed samples discloses the very different looking diagram of Fig. 162. On account of the very high initial permeabilities reached, this diagram is drawn to a smaller vertical scale than the one just considered. The rapid cooling of the air-quenched samples produces particularly marked changes in a portion of the iron-nickel range. In this binary series the values are about the same as those of the annealed samples up to a nickel content of about 45 per cent. From here on, as the nickel is increased, instead of a distinct valley and a secondary rise noticed in Fig. 161, there is a continued rise reaching a peak value of about 8,000 at the composition of 78½ per cent nickel, 21½ per cent iron. Air quenching is thus shown to have increased the initial permeability to more than four times that of the annealed sample of the same composition.

Other alloys, however, such as those of the iron-cobalt series, are affected in the opposite way by air quenching. The hump in the neighborhood of 50 per cent iron, which, in the annealed samples, showed an initial permeability of about 600 has fallen, in the air-quenched series, to a value of about 400.

Coming now to the maximum instead of the initial permeabilities for annealed alloys, Fig. 163 may be compared with Fig. 161. The forms of the two diagrams are similar in a very general way, though the maximum values shown in Fig. 163 are, of course, very much greater. The most pronounced difference is the striking decrease in the maximum permeability of iron as small

percentages of either nickel or cobalt are added. As much as 10 per cent of either reduces the maximum permeability of Armco iron by a factor of more than 6.

The solid diagram of maximum permeabilities for air-quenched alloys is said to resemble generally the one (Fig. 162) for initial

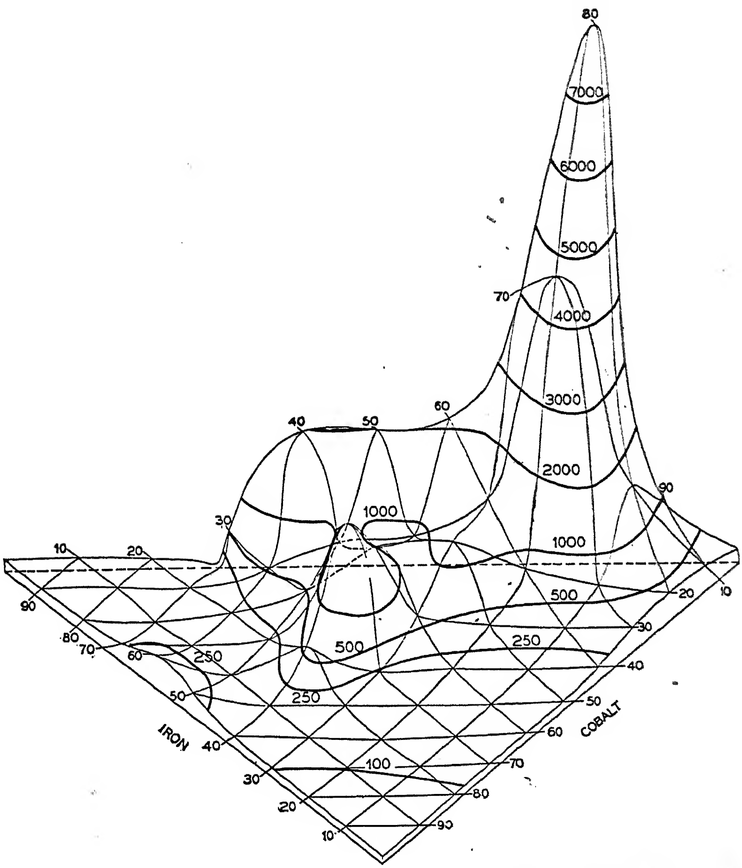


FIG. 162.—Initial permeabilities for air-quenched alloys (*Courtesy of Bell Telephone Laboratories.*)

permeabilities of those alloys. The maximum permeabilities for these rapidly cooled alloys are, however, enormously higher, reaching a peak of 120,000 for the 78.5 per cent nickel 21.5 per cent iron air-quenched alloy. Even this is not the maximum attained for the composition, for, in commenting on the great influence that the rate of cooling has on the permeability, Elmen

mentions having secured initial values of 13,000 and maximum values of 400,000 from other castings of the same composition.

The name "permalloy" has been applied to that series of iron-nickel alloys which is remarkable for high permeability at low magnetizing force. In referring to the various compositions

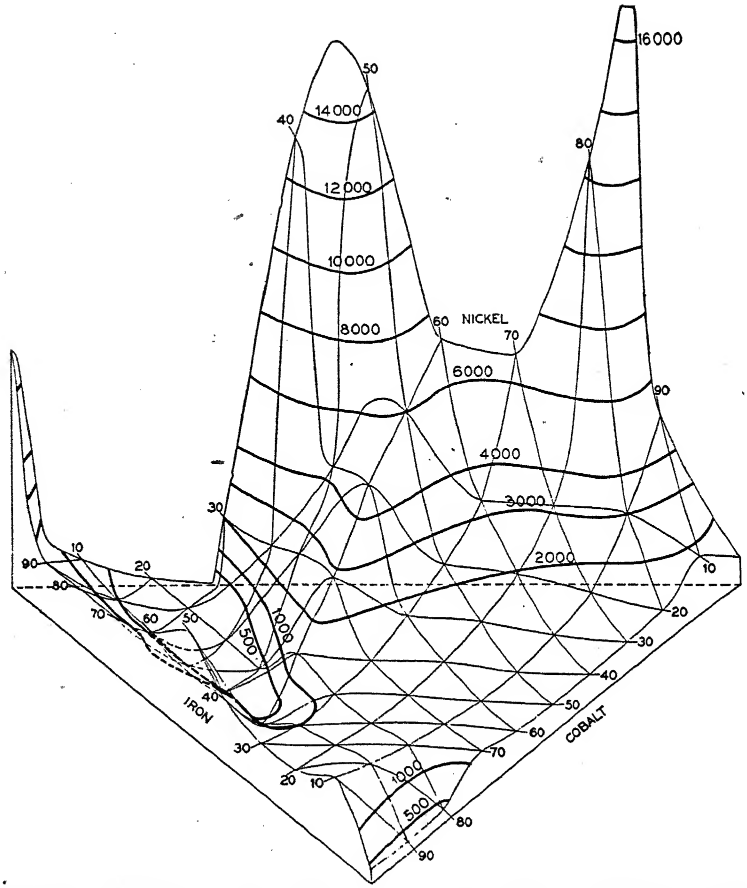


FIG. 163.—Maximum permeabilities for annealed alloys. (*Courtesy of Bell Telephone Laboratories.*)

it has been found convenient to use the nickel content percentage as a distinguishing prefix. Thus 78.5 per cent permalloy is an alloy composed of 78.5 per cent nickel and 21.5 per cent iron.

Figure 164 gives the magnetization or *B-H* curves for Armco iron, for air-quenched and for annealed 78.5 per cent permalloy. The insert in this figure shows the lower sections of these curves

on a larger scale, so as to bring out more clearly the flux densities at very low magnetizing forces.

In Fig. 165 the corresponding permeability curves are plotted. The enormous difference between the values for air-quenched and annealed permalloy is to be noted, as is also the high initial values of both in comparison with the best grade of iron. This figure in connection with Fig. 158 affords a basis of comparison of the permeabilities of Armco iron and silicon steel.

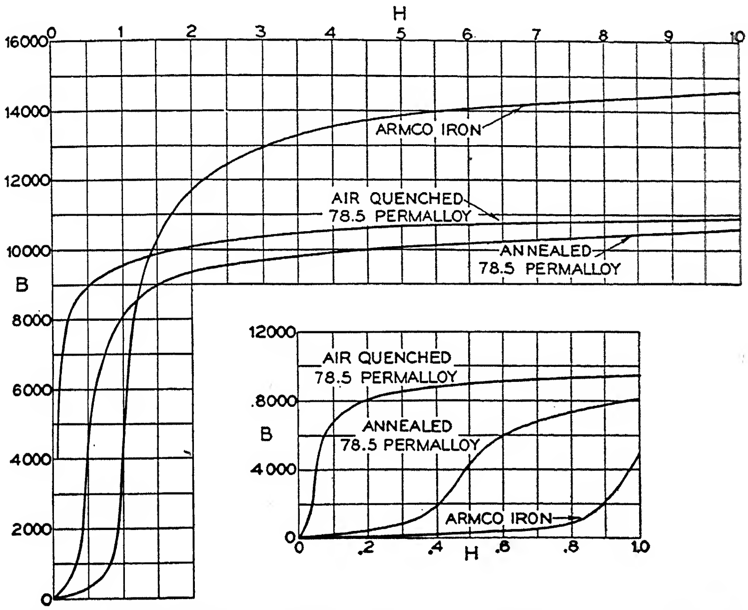


FIG. 164.—Magnetization curves for 78.5 permalloy and Aruco iron. (*Courtesy of Bell Telephone Laboratories.*)

The quest for alloys of high permeabilities at low magnetizing forces did not end with the production of permalloy, or with the consideration of the three magnetic metals alone. The effect of adding non-magnetic elements to various alloys of the magnetic metals was studied with further remarkable results. As an example, the addition of 3.5 per cent of the non-magnetic metal molybdenum to 78.5 permalloy (designated "3.5–78.5 mo-permalloy") resulted in a composition having very much higher initial permeability than the straight 78.5 permalloy. Figure 166 shows the permeability curves for this mo-permalloy, 78.5 permalloy and Armco iron at very low magnetizing forces up to 0.03 gauss. In general, the addition of some of the non-magnetic

metals to permalloys as well as to various ternary alloys of iron, nickel and cobalt made the alloys less sensitive to heat treatment and tended to increase their resistivity.

The great interest that has been attracted by the remarkable magnetic properties of the permalloys has, perhaps, caused the impression that the real achievements in Elmen's work all lie in

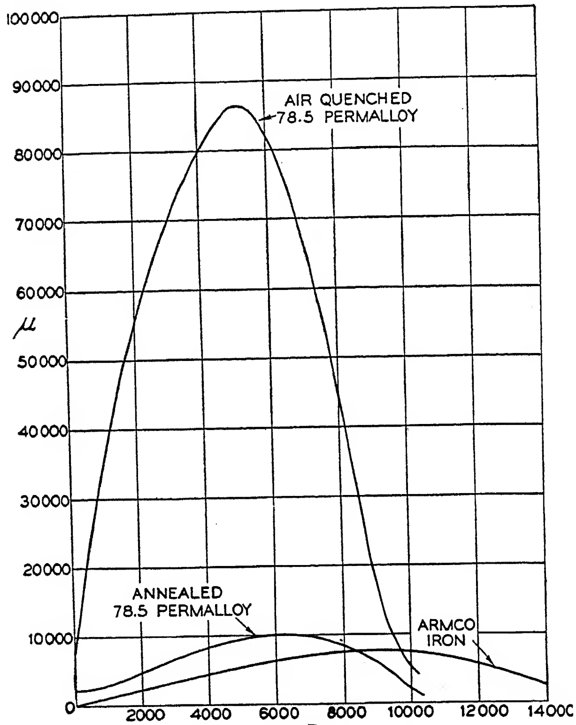


FIG. 165.—Permeability curves for 78.5 permalloy and Armco iron. (*Courtesy of Bell Telephone Laboratories.*)

the field of low magnetic intensities. This is true, perhaps, of the nickel-iron alloys, where the most remarkable magnetic properties are found at low magnetizing forces, but it is not true of the cobalt-iron alloys which deserve attention because of their behavior at very high flux densities.

Figure 167 is a solid diagram showing the *intrinsic flux* or *intrinsic induction* for all of the iron-nickel-cobalt annealed alloys at magnetizing forces H of 50 gauss. Intrinsic flux or induction differs slightly from total flux or induction. It is

the part of the induction that is contributed by the magnetic material. In other words, it is B minus H. This diagram of Fig. 167 has been turned around from the position occupied by the solid diagrams of Figs. 161, 162 and 163. Here the left-hand corner represents 100 per cent cobalt, the right-hand corner

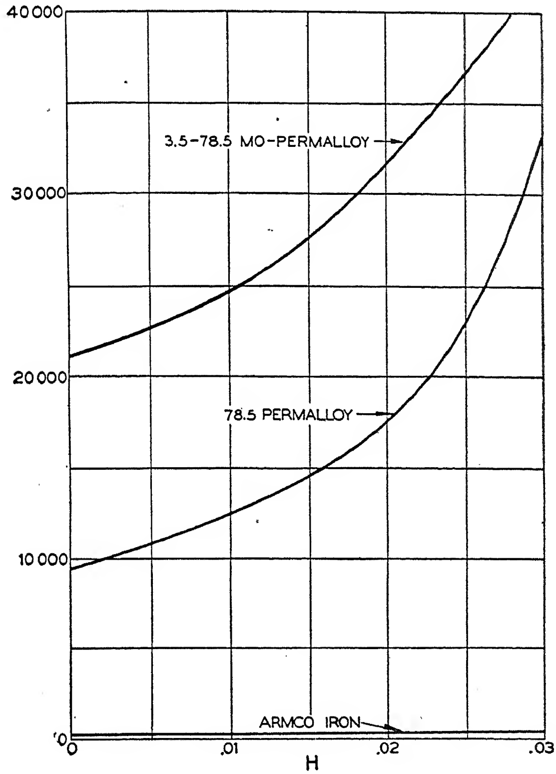


Fig. 166.—Permeability curves for two of the permalloys and Armco iron at low-magnetizing forces. (*Courtesy of Bell Telephone Laboratories.*)

100 per cent iron, and the front corner 100 per cent nickel. The rear face is, therefore, the iron-cobalt face.

We see from the profile of the iron-cobalt face that while 100 per cent iron shows, for this magnetizing force of 50 gauss, a flux density of approximately 16,000 lines per square centimeter and 100 per cent cobalt a flux density of less than 10,000 lines per square centimeter, there is a long range of the binary alloys that substantially exceed either of these figures. A

peak of nearly 23,000 is reached at the 50 per cent iron-50 per cent cobalt alloy. The same general characteristics are indicated in the solid diagram, not here shown, for the high magnetizing force of H = 1,500 gauss.

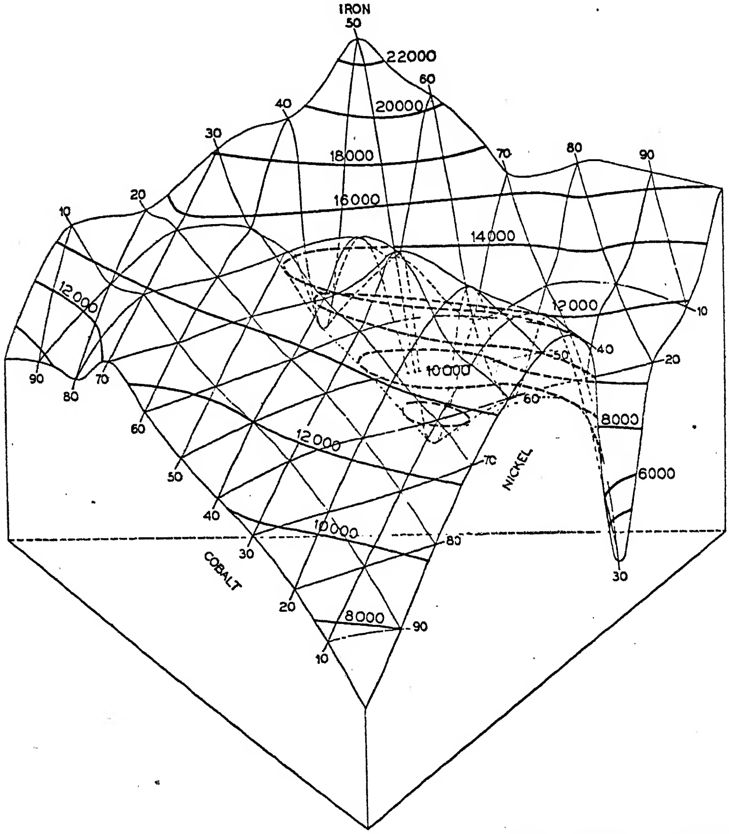


FIG. 167.—Intrinsic inductions for annealed alloys at H = 50. (*Courtesy of Bell Telephone Laboratories.*)

Examining this particular 50-50 alloy of cobalt and iron throughout a wide range of magnetizing forces has resulted in the curves of Fig. 168. In this figure the scale of magnetizing force H of the magnetization curves is plotted at the top, and that of permeabilities μ for the permeability curves at the bottom. Common to both curves, the scale of *intrinsic* flux or *intrinsic* induction, $B - H$, is plotted at the left.

From the curves of Fig. 168 it will be seen that, unlike permalloy, this 50-50 iron-cobalt alloy is not remarkable for its high permeability at low magnetizing forces, although its inital permeability is about 600 against 250 for Armco iron. Neither is it remarkable for its maximum permeability, which is about 4,000 as against 7,000 for Armco iron. It is at inductions above 13,000 lines per square centimeter that this alloy begins to show

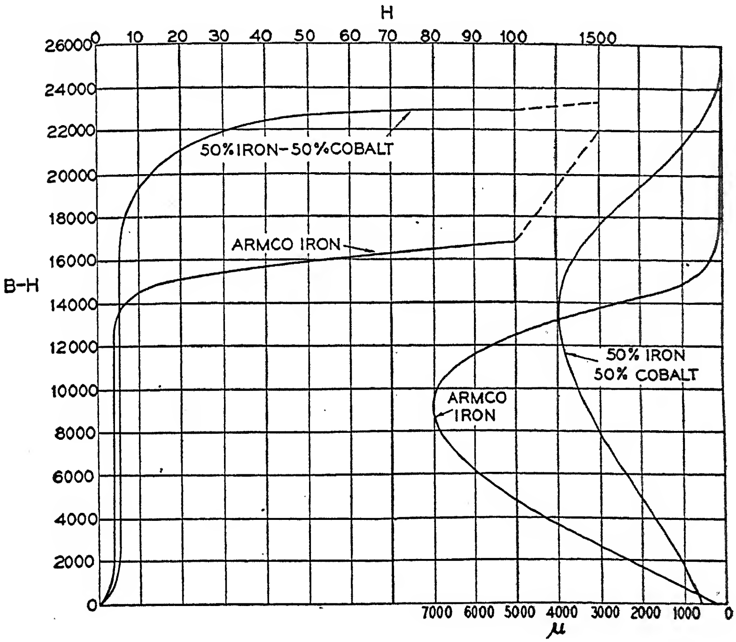


FIG. 168.—Magnetization and permeability curves for 50-50 iron-cobalt alloy and for Armco iron. (*Courtesy of Bell Telephone Laboratories.*)

its outstanding magnetic properties. Above this flux density and for inductions up to about 25,000, its permeability is far superior to Armco iron or Silicon steel. For a magnetizing force of 100 gauss the intrinsic induction of this alloy is about 23,000 lines per square centimeter as against a value of 17,000 for Armco iron. We have then, in this alloy, a magnetic material of much more uniform permeability than the best grade of iron throughout the entire working range and of markedly higher permeability for flux densities far beyond the saturation point of iron.

The hysteresis losses of the entire range of iron-nickel-cobalt alloys have also been studied for different maximum flux densities and for different heat treatments. Figure 169 is a solid diagram of hysteresis losses for annealed alloys of all compositions for a maximum flux density of 5,000 lines per square centimeter, plus and minus. Here the vertical heights above the plane of the base represent the hysteresis losses in ergs per cubic centimeter per cycle. The comparatively high losses for compositions in which iron predominates are to be noted. It is significant that the lowest energy losses of all the compositions occur in

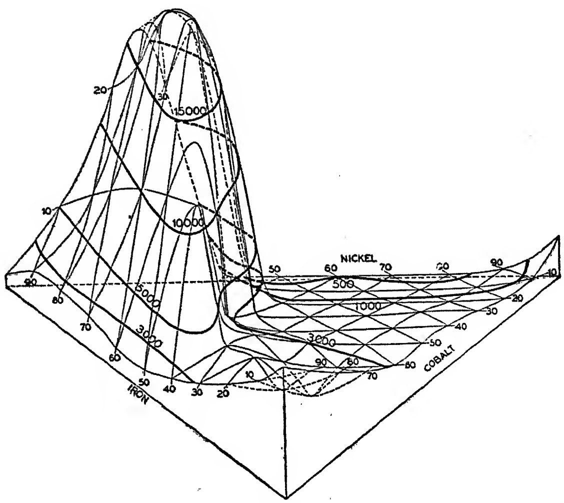


FIG. 169.—Hysteresis losses per cubic centimeter per cycle for maximum flux density of $B = 5{,}000$ for annealed alloys. (*Courtesy of Bell Telephone Laboratories.*)

the neighborhood of the 78.5 permalloy; and, also, that in the iron-cobalt series, the 50-50 iron-cobalt alloy has the lowest loss.

A group of ternary alloys characterized by constant permeability and extremely low hysteresis loss have been termed *perminvars.* Typical of these may be taken the alloy comprising 45 per cent nickel, 25 per cent cobalt and 30 per cent iron. The permeability curves for this alloy in its baked, annealed and air-quenched conditions respectively are shown in Fig. 170. In the insert diagram in the upper right-hand corner of this figure

the lower portions of these same curves have been drawn to larger scale to illustrate more clearly the permeability at low

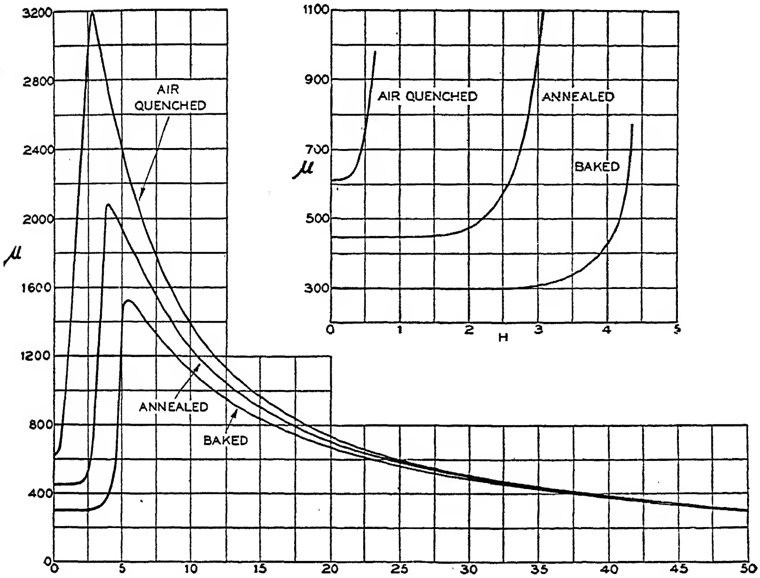


FIG. 170.—Permeability curves for perminvar—45 per cent nickel, 25 per cent cobalt, and 30 per cent iron. (*Courtesy of Bell Telephone Laboratories.*)

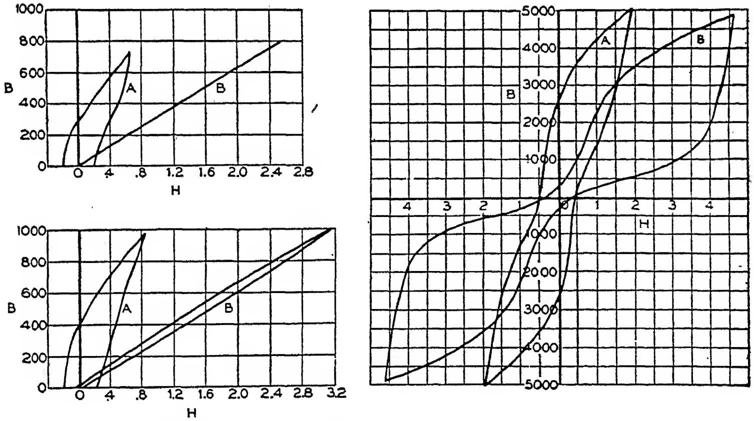


FIG. 171.—Hysteresis characteristics for perminvar. *A, air-quenched. B, baked.* (*Courtesy of Bell Telephone Laboratories.*)

magnetizing forces. The remarkable constancy of permeability, particularly for the baked sample, is to be noted, its value

remaining unchanged at 300 for magnetizing forces from zero to $2\frac{3}{4}$ gauss.

The hysteresis loops for three different maximum flux densities, $B = 750$, $B = 1,000$ and $B = 5,000$, are shown in Fig. 171 for both air-quenched and baked samples of this same 45 per cent nickel, 25 per cent cobalt and 30 per cent iron perminvar composition. The remarkable behavior of the baked sample (curve B in the three diagrams) is to be noted. At the lower density of $B = 750$ the ascending side of the loop coincides with the descending, leaving no measurable area in the loop and indicating correspondingly low energy losses. For the next higher density, $B = 1,000$, the loop for the baked alloy begins to have measurable area, but is still considerably smaller than that of the annealed sample. This condition is reversed for the maximum flux density of $B = 5,000$, where the loops indicate a higher loss for the baked than for the annealed alloys. A constriction in the width of the loops near the point of origin is a marked characteristic of the baked alloy of this composition. The fact that both the magnetization and demagnetization curves approach closely to the origin in passing through the lower flux densities indicate exceedingly small retentivity and coercive force for the baked samples, even where the maximum density of the cycle is as high as 5,000 lines per square centimeter.

Improvement in magnetic materials during recent years has not been confined to the mere composition and heat treatment of the material itself. While Elmen and his associates in the Bell Telephone Laboratories have been producing the magnetic alloys, some of which have been just referred to, others, principally in the same laboratories, have been working on better means of subdividing the mass of magnetic core material so as to secure lower core losses and greater stability of magnetic behavior under widely varying conditions.

The practice of subdividing the mass of magnetic material in cores and other parts of magnetic circuits in order to reduce eddy currents and to secure other desirable results has long been followed. Instead of employing solid metal, cores for many years have been made of bundles of iron wire or of flat iron strips, as in induction coils, and of piles of flat sheets of iron, as in transformers and in the armatures of dynamos and motors. Within recent years, however, this idea of the subdivision of cores has been carried much further in the production of the so-called "dust core."

Somewhat over ten years ago and before permalloy was available, cores of compressed powdered iron had been extensively employed in loading coils used for improving the transmission qualities of telephone lines. These cores were made by grinding up electrolytic iron into a very fine dust. The individual grains of iron were then insulated with an oxide coating and shellac and then formed into rings of the desired shape under very high pressures.

Since the advent of permalloy, it has replaced iron in the production of these cores. The method of making compressed powdered permalloy is briefly as follows:[1]

The alloy, cast in such manner as to make it brittle, is ground into powder, annealed, and again ground. The resulting very fine powder is of silver white appearance, the size of the particles varying widely.

The particles are then given an insulating coating of refractory material which will not lose its insulating qualities under the heat treatment to which the molded cores must be subjected. The insulated powder is then formed into cores in dies under pressure of about 100 tons per square inch. The degree of pressure employed, up to about 100 tons, has marked effect on the resulting permeability, the permeability increasing rapidly with increased pressure up to about that point.

The heat treatment of the compressed cores is a delicate process. Subjecting them to too high temperatures or for too long a time will break down the insulation and lower the resistivity. On the other hand, too low temperatures will not develop the desired magnetic characteristics, particularly as to high permeability and low hysteresis loss.

The resulting compressed powdered permalloy has a density nearly as great as that of the solid alloy. It has sufficient mechanical strength for the practical purposes of subsequent manufacturing operations.

The resistivity of the material varies considerably. Typical values are from 1 to 20 ohms per cubic centimeter, and it is stated that this may vary over a considerable range without having significant effect on the magnetic properties. Variation below this range, however, is an indication of incomplete insula-

[1] Shackelton, W. J., and I. C. Barber, Compressed Powdered Permalloy, Manufacture and Magnetic Properties, American Institute of Electrical Engineers, *Transactions*, vol. 47, pp. 429–436, April, 1928.

tion of the particles and of higher than normal eddy-current losses.

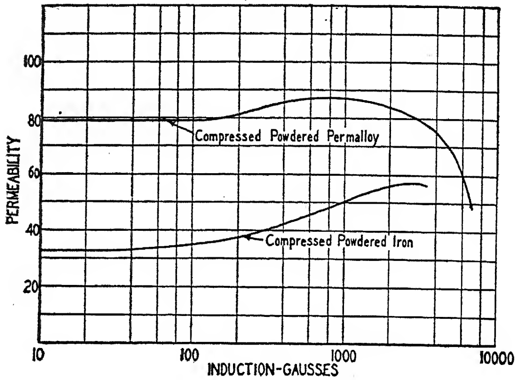


FIG. 172.—Permeability-induction characteristics of compressed powdered permalloy and compressed powdered iron. (*Courtesy of Bell Telephone Laboratories.*)

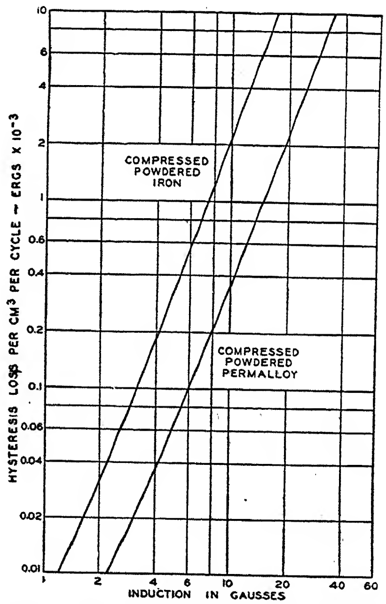


FIG. 173.—Hysteresis-loss characteristics of compressed powdered permalloy and compressed powdered iron. (*Courtesy of Bell Telephone Laboratories.*)

The permeabilities of compressed powdered permalloy and of compressed powdered iron for different values of flux density are plotted, respectively, in the two curves of Fig. 172, values

of B being plotted on the logarithmic scale. It is to be noted that the permalloy curve shows: (1) larger initial permeability, (2) a wider range of induction over which the permeability is practically constant and (3) smaller percentage change of permeability over a substantially wider range of flux density.

The eddy-current and hysteresis losses of the compressed powdered permalloy are much smaller than those of any other material of comparable constancy of permeability. Figure 173 gives the hysteresis loss in ergs per cubic centimeter per cycle for various flux densities. It is evident that for a given hysteresis loss powdered permalloy is capable of operating at much higher flux densities than powdered iron.

It must be remembered that this finely divided magnetic material has been developed to meet the magnetic requirements of one highly specialized use. It seems likely that material of widely differing magnetic qualities may eventually be developed for other special applications.

In the dust core we have a magnetic material capable of being molded into simple shapes. There has always been something of a mechanical problem involved in interlinking a closed magnetic circuit with a closed coil of wire. Either the magnetic "ring" had to be split to get the winding on or special winding methods employed. The writer suggests that, eventually, a better solution may be developed by molding the closed core around and through the winding, thus interlinking the two without either splitting the ring or employing special winding machinery. While obviously difficult, this problem appears worthwhile.

The benefits to electrical communication which have already resulted from the quest for more perfect magnetic materials may be illustrated by the use of compressed powdered permalloy as the core material of loading coils. Iron-dust cores permitted the construction of loading coils that were much smaller and cheaper than the older coils having iron wire cores. Permalloy dust cores enabled a further large saving in size and cost to be made, so that the modern coil is only a fraction of the size and cost of the older ones. The saving in size alone has not only made for cheaper coils but also has permitted important economies to be made in both aerial and underground structures required for their support or housing. But more important than the economies due to size or first cost, the new coils are more efficient and more stable. This subject will be more fully

treated in a subsequent chapter; the point here is that this particular development of a more perfect magnetic material for a specific purpose has resulted not only in better telephone lines but also in a saving of many millions of dollars per annum.

From a broader aspect it appears that magnetic materials are not only to serve their old functions more effectively but also actually to invade new fields. One revolutionary accomplishment has already been made in a field heretofore not employing magnetic materials at all—that of the conductors in submarine telegraph cables. By wrapping the conductor with a tape of permalloy, what is called "continuous loading" is accomplished. The inductance of each small length of conductor is thus made to offset the ill effects of the capacity of that length. The principle of continuous loading has long been known, but until permalloy was produced there was no known magnetic material having anywhere near high enough permeability to accomplish the result economically on long submarine cables. The first permalloy loaded cable was between New York and the Azores, a distance of 2,300 miles. This proved to have five times the message-carrying capacity of cables of the old type. Since this success, many thousands of miles of cable of this type have been laid.

The similar treatment of the conductors of telephone cables gives the only reasonable promise that has yet been made of transmitting the human voice across the Atlantic by cable, a feat that until recently has been considered quite impossible of commercial achievement. Such a cable is now in process of construction, and is made possible solely by the availability of a new magnetic material.

PART III

ELEMENTS OF APPARATUS

The word element is here used in the sense of one of the essential parts of which a thing is composed. Of such elements of telephone apparatus there are a number, such as wire, coils, cords, resistors, condensers and contacts, which are so generally used throughout the whole realm of telephone equipment as to warrant their being separately discussed without regard to the particular piece of apparatus of which they may eventually become component parts.

CHAPTER XI

WIRES FOR EQUIPMENT USE

It is the intent of this chapter to deal with those wires more intimately related to the making and wiring of *telephone equipment* as distinguished from those principally used in the construction of *telephone lines*. The distinction may not be too sharply drawn, for, obviously, there is some overlapping. We might refer to the wires of this chapter as "interior" or "inside" wires to differentiate from the "outside" wires mostly used in out-of-door line construction. This, however, might lead to some confusion, if for no other reason, because the names "interior" and "inside" as referring to wires have come to have a much narrower meaning. Much that is here said, however, will apply also to wires used for line-construction work, particularly in so far as conductivity and size of wire are concerned.

Conductor Characteristics.—For many purposes the most important electrical characteristic of a conductor is its conductivity, or its resistance, which is the reciprocal of its conductivity. Table I shows the resistance of various metal conductors in terms of pure copper as given by the American Telephone and Telegraph Company. It is always to be remembered that there are likely to be some variations in the figures given in different tables for the same metal. These are usually due to the variations in resistance caused by impurities in the metal or by the different physical states, such as of hardness, in which the same metal may exist. An example of such differences is found in this table in the various coppers, irons and steels.

Because of its high conductivity and great ductility, its lasting qualities under atmospheric conditions and its relatively low cost, copper is by far the most important metal for electrical wires. Of all the metals, silver and gold alone approach it in conductivity, silver being a slightly better and gold a somewhat poorer conductor. The high costs of silver and gold preclude their large use.

Table I.—Resistance of Various Metal Conductors
Compared to pure copper of same length and cross-section

Kind of Metal	Times Resistance of Pure Copper
Silver	0.975
Pure copper	Unity
Annealed copper	1.032
Hard drawn copper	1.067
Gold	1.325
Aluminum	1.815
Magnesium	2.95
Zinc	3.76
Tungsten	4.54
Brass	5.52
Tin	6.79
Iron, commercial	7.02
Nickel	7.54
Platinum	7.96
Tantalum	9.43
Soft steel	10.6
Lead	12.6
German silver	19.5
Hard steel	29.6
Mercury	61.0
Cast iron	64.0

Aluminum as a conductor ranks fourth among the metals, with somewhat more than one-half the conductivity of copper for the same size conductor.

The best grades of iron wire have only about one-sixth the conductivity of copper, size for size, and the cheapness of iron in comparison with copper has led to its use for line wire to a considerable extent. Its perishable qualities, however, are strongly against it, even under conditions where its low conductivity may be tolerated.

Except in those cases where resistivity instead of conductivity is the quality desired, copper, for the foregoing principal reasons, is the outstanding metal for electrical conductors.

The diameter of wire for electrical purposes is usually expressed in terms of some so-called "standard wire gage," and there are, unfortunately, a number of such. Most of the different gages have been brought into existence by wire manufacturers or by other organizations attempting to sanction officially some orderly arrangement of gage numbers. In these gages the sizes

of wires are referred to by numbers, and, in most cases, the smaller numbers refer to the larger wires.

Another way of designating wire sizes is to ignore fixed series of gage numbers and refer merely to the diameter in thousandths of an inch or in mils, as thousandths of an inch are called. Again we may refer to the size of a wire in terms of its cross-sectional area expressed in "circular mils." A circular mil is the area of a circle the diameter of which is one mil. This is better than expressing the area in square inches, or in square thousandths of an inch, because the area in circular mils is obtained by simply squaring the diameter of the wire in mils. This simple relationship is true, of course, because the areas of two circles are to each other as the squares of their diameters. To reduce circular mils to square inches, multiply by 0.7854 and divide by 1,000,000.

It is not feasible, however, to ignore a fixed series of gage numbers. In the first place, it is a great convenience to be able to refer merely by number to a comparatively few recognized sizes whose diameters, areas, resistances and weights are all definitely associated with the respective numbers. In the second place, it is of great advantage for manufacturers to be able to standardize on a fixed series of sizes rather than to be required to make and carry in stock an unlimited number of sizes.

Fortunately, so far as copper wire is concerned, but a single wire gage remains in general use. This is the Brown and Sharpe gage (B. & S. G.) also called the American Wire gage (A. W. G.). As one sometimes finds reference to other wire gages, however, even in connection with copper wire, and as some of them are still used in designating the thickness of other kinds of wire and of sheet metal, Table II is here given for general reference.

The Brown and Sharpe gage, however, is the pertinent one to consider in practically all copper-wire computations. It was devised by Mr. J. R. Brown, in 1857, with particular regard to wire-drawing practice. The various characteristics of diameter, area, weight and resistance for the different numbers of this gage are given in Table III. In this table the resistances are those for commercial copper which is 99.99 per cent pure, as required by the American Society for Testing Materials.

It is a characteristic of the B. & S. gage that the successive diameters vary in geometrical progression, the common ratio between the diameters of successive gage numbers being 1.1229.

TABLE II.—COMPARISON OF WIRE GAGES

Diameters in inches

Gage No.	Brown & Sharpe gage (A.W.G.) (B. & S.)	Birmingham wire gage (B.W.G.)	Old English wire gage (London)	New British standard wire gage (N.B.S.G.)	Steel wire gage[1]	Stubs' steel wire gage	American Screw Co.'s screw wire gage
4-0	0.4600	0.454	0.454	0.400	0.3938		
3-0	0.4096	0.425	0.425	0.372	0.3625	.....	0.0315
2-0	0.3648	0.380	0.380	0.348	0.3310	.....	0.0447
0	0.3249	0.340	0.340	0.324	0.3065	.....	0.0578
1	0.2893	0.300	0.300	0.300	0.2830	0.227	0.0710
2	0.2576	0.284	0.284	0.276	0.2625	0.219	0.0842
3	0.2294	0.259	0.259	0.252	0.2437	0.212	0.0973
4	0.2043	0.238	0.238	0.232	0.2253	0.207	0.1105
5	0.1819	0.220	0.220	0.212	0.2070	0.204	0.1236
6	0.1620	0.203	0.203	0.192	0.1920	0.201	0.1368
7	0.1443	0.180	0.180	0.176	0.1770	0.199	0.1500
8	0.1285	0.165	0.165	0.160	0.1620	0.197	0.1631
9	0.1144	0.148	0.148	0.144	0.1483	0.194	0.1763
10	0.1019	0.134	0.134	0.128	0.1350	0.191	0.1894
11	0.09074	0.120	0.120	0.116	0.1205	0.188	0.2026
12	0.08081	0.109	0.109	0.104	0.1055	0.185	0.2158
13	0.07196	0.095	0.095	0.092	0.0915	0.182	0.2289
14	0.06408	0.083	0.083	0.080	0.0800	0.180	0.2421
15	0.05707	0.072	0.072	0.072	0.0720	0.178	0.2552
16	0.05082	0.065	0.065	0.064	0.0625	0.175	0.2684
17	0.04526	0.058	0.058	0.056	0.0540	0.172	0.2816
18	0.04030	0.049	0.049	0.048	0.0475	0.168	0.2947
19	0.03589	0.042	0.040	0.040	0.0410	0.164	0.3079
20	0.03196	0.035	0.035	0.036	0.0348	0.161	0.3210
21	0.02846	0.032	0.0315	0.032	0.0317	0.157	0.3342
22	0.02535	0.028	0.0295	0.028	0.0286	0.155	0.3474
23	0.02257	0.025	0.0270	0.024	0.0258	0.153	0.3605
24	0.02010	0.022	0.0250	0.022	0.0230	0.151	0.3737
25	0.01790	0.020	0.0230	0.020	0.0204	0.148	0.3868
26	0.01594	0.018	0.0205	0.018	0.0181	0.146	0.4000
27	0.01420	0.016	0.01875	0.0164	0.0173	0.143	0.4132
28	0.01264	0.014	0.01650	0.0148	0.0162	0.139	0.4263
29	0.01126	0.013	0.01550	0.0136	0.0150	0.134	0.4395
30	0.01003	0.012	0.01375	0.0124	0.0140	0.127	0.4526
31	0.008928	0.010	0.01225	0.0116	0.0132	0.120	0.4658
32	0.007950	0.009	0.01125	0.0108	0.0128	0.115	0.4790
33	0.007080	0.008	0.01025	0.0100	0.0118	0.112	0.4921
34	0.006305	0.007	0.00950	0.0092	0.0104	0.110	0.5053
35	0.005615	0.005	0.00900	0.0084	0.0095	0.108	0.5184
36	0.005000	0.004	0.00750	0.0076	0.0090	0.106	0.5316
37	0.004453	.....	0.00650	0.0068	0.0085	0.103	0.5448
38	0.003965	.....	0.00575	0.0060	0.0080	0.101	0.5579
39	0.003531	.....	0.00500	0.0052	0.0075	0.099	0.5711
40	0.003145	.....	0.00450	0.0048	0.0070	0.097	0.5842

[1] The steel wire gage is the same gage which has been known by various names: "Washburn and Moen," "Roebling," "American Steel and Wire Co.'s." Its abbreviation should be written "Steel W. G." to distinguish it from "S.W.G.," a common abbreviation for the British standard wire gage.

Table III.—Bare Copper Wire Tables

Brown and Sharpe gage

Giving measurements at 68° F. (20° C.) with specific gravity of 8.89

A. W. G., B. & S. gage	Diameter, inches	Area, circular mils	Weight, pounds per 1,000 ft.	Length, feet per pound	Resistance	
					Ohms per 1,000 ft.	Ohms per pound
0000	0.4600	211,600.	640.5	1.561	0.04901	0.00007652
000	0.4096	167,800.	507.9	1.969	0.06180	0.0001217
00	0.3648	133,100.	402.8	2.483	0.07793	0.0001935
0	0.3249	105,500.	319.5	3.130	0.09827	0.0003076
1	0.2893	83,690.	253.3	3.948	0.1239	0.0004891
2	0.2576	66,370.	200.9	4.978	0.1563	0.0007778
3	0.2294	52,630.	159.3	6.276	0.1970	0.001237
4	0.2043	41,740.	126.4	7.911	0.2485	0.001966
5	0.1819	33,100.	100.2	9.980	0.3133	0.003127
6	0.1620	26,250.	79.46	12.58	0.3951	0.004972
7	0.1443	20,820.	63.02	15.87	0.4982	0.007905
8	0.1285	16,510.	49.98	20.01	0.6282	0.01257
9	0.1144	13,090.	39.63	25.23	0.7921	0.01999
10	0.1019	10,380.	31.43	31.82	0.9989	0.03178
11	0.09074	8,234.	24.92	40.13	1.260	0.05053
12	0.08081	6,530.	19.77	50.58	1.588	0.08035
13	0.07196	5,178.	15.68	63.77	2.003	0.1278
14	0.06408	4,107.	12.43	80.45	2.525	0.2032
15	0.05707	3,257.	9.858	101.4	3.184	0.3230
16	0.05082	2,583.	7.818	127.9	4.016	0.5136
17	0.04526	2,048.	6.200	161.3	5.064	0.8167
18	0.04030	1,624.	4.917	203.4	6.385	1.299
19	0.03589	1,288.	3.899	256.5	8.051	2.065
20	0.03196	1,022.	3.092	323.4	10.15	3.283
21	0.02846	810.1	2.452	407.8	12.80	5.221
22	0.02535	642.4	1.945	514.1	16.14	8.301
23	0.02257	509.5	1.542	648.5	20.36	13.20
24	0.02010	404.0	1.223	817.7	25.67	20.99
25	0.01790	320.4	0.9699	1,031.	32.37	33.37
26	0.01594	254.1	0.7692	1,300.	40.81	53.06
27	0.01420	201.5	0.6100	1,639.	51.47	84.37
28	0.01264	159.8	0.4837	2,067.	64.90	134.2
29	0.01126	126.7	0.3836	2,606.	81.83	213.3
30	0.01003	100.5	0.3042	3,287.	103.2	329.2
31	0.008928	79.70	0.2413	4,144.	130.1	539.3
32	0.007950	63.21	0.1913	5,227.	164.1	857.6
33	0.007080	50.13	0.1517	6,591.	206.9	1364.
34	0.006305	39.75	0.1203	8,312.	260.9	2168.
35	0.005615	31.52	0.09542	10,480.	329.0	3448.
36	0.005000	25.00	0.07568	13,213.	414.8	5482.
37	0.004453	19.83	0.0601	16,664.	523.1	8717.
38	0.003965	15.72	0.04759	21,012.	659.6	13860.
39	0.003531	12.47	0.03774	26,497.	831.8	22040.
40	0.003145	9.888	0.02990	33,411.	1,049.	35040.
(41)*	0.00275	7.5625	0.02289	43,700.	1,370.	59,900.
(42)*	0.00250	6.2500	0.01892	52,800.	1,660.	87,700.
(43)*	0.00225	5.0625	0.01532	65,300.	2,050.	133,700.
(44)*	0.00200	4.0000	0.01211	82,600.	2,600.	214,000.
(45)*	0.00175	3.0625	0.00927	107,900.	3,390.	365,200.
(46)*	0.00150	2.2500	0.00681	146,800.	4,610.	676,800.

* B. & S. gage numbers for the sizes smaller than No. 40 are often used but are not yet fully recognized. It is best to specify these sizes by their diameters.

As a consequence, the cross-sectional areas, weights and resistance of the corresponding wires also constitute a geometrical progression with a common ratio for successive gage numbers of 1.261.

As 1.261 is very close to the cube root of 2, it follows that an increase of 3 in gage numbers will double the resistance and halve the cross-sectional area and weight per unit length. By the same token, as 1.1229 is about the sixth root of 2. an increase of six gage numbers will halve the diameter.

Again, as 10 is close to the tenth power of 1.261, it follows that an increase of 10 in gage numbers will multiply the resistance by 10, and divide the area and weight by the same number.

To state these relationships more succinctly, the diameter is doubled or halved every sixth number as one proceeds up or down the scale of numbers. The area, weight and resistance are doubled or halved every third number, and are multiplied or divided by ten every tenth number.

If one will carry in mind these simple ratios characterizing the B. & S. gage, he may, without a table before him, approximate the diameter, area, weight or resistance of any size of copper wire, provided he has a suitable starting point from which to figure. Such a starting point is found in No. 10 wire, easily remembered because of the recurrence of the number 10 or its multiples. Within reasonable accuracy, the diameter of No. 10 wire is $\frac{1}{10}$ inch, the area 10,000 circular mils and the resistance of 1,000 feet is 1 ohm. If one can also remember that the weight of 1,000 feet is 31.43 pounds, he has, within reach, a fair approximation to all the information as to bare copper wire that is contained in such a compilation as that of Table III.

Wire Insulating Materials.—For the construction or wiring of telephone equipment, bare copper wire is seldom used. It is generally provided with a covering of insulating material for purposes of electrical separation and, in lesser degree, mechanical protection. Various insulating materials are used for this purpose, and they are applied to the wire in different ways. This has resulted in many kinds of insulated wires named according to the character of their insulation or the uses to which they are to be put. Thus we have "magnet wire" used for winding electromagnets, "switchboard wire" for connecting together the various parts of a telephone switchboard, "jumper wire" for making cross-connections on distributing frames,

"inside telephone wire" for wiring the inside of subscribers' premises, "rubber-covered wire" where especially high insulation requirements are to be met, and so on. The materials used for the insulation of such wires are the so-called "textiles," of which the principal ones are silk and cotton; an enamel composed of gums and oils; and rubber. The various characteristics of these insulating materials may be briefly referred to.

A short article by Howard H. Glenn[1] gives much information concerning textile insulating materials, some of which is briefly summarized here. Of real silk, there are two general classes, cultivated and wild. Cultivated silk is that obtained from the kind of silkworm which feeds on mulberry leaves and is raised under human care. Wild silk is produced by other varieties of silkworms not amenable to such culture, and which feed on oak, fig and castor-oil plant leaves. This wild silk is the so-called "tussah silk" of commerce.

Tussah silk is the cheaper of the two varieties. Its fibers are coarser and it is more difficult to dye. Its electrical properties, however, are good, and it is widely used as an insulating material for wires and cords, particularly where it is to be covered by a layer of some other kind of insulating material.

The finer grades of cultivated silks, while much more expensive, are largely used for wire insulation in those cases where economy of space is important and, therefore, where very thin layers of covering are required. Their long and very fine fibers render them superior to all other textiles under these conditions.

A cheaper form of cultivated silk, however, compares favorably in cost with tussah silk. This is called "spun silk" and is made from the shorter silk fibers of cultivated silk, resulting from defective cocoons, and also from the wastage in the manufacture of long-fiber silk.

The facility with which a textile takes dye is often important. In this respect, where coloring is necessary either for the purpose of identification or for that of attractive appearance, cultivated silk is far superior to tussah. On the other hand, there are many cases where the natural brownish color of tussah silk is in no wise objectionable, as where it is used as an inner layer of insulation, covered by an outer layer of some material which will take color as desired.

[1] Textiles for Insulation in Telephone Equipment, *Bell Laboratories Record*, p. 53, April, 1926.

For various reasons none of the different forms of artifical silk have found their way into large use as substitutes for either of these forms of real silk.

Cotton, also, is one of the most valuable of the textile insulating materials. In insulating value it ranks high, but inferior to silk. It affords an excellent covering mechanically. Its value is further enhanced by the fact that it takes dye readily. It cannot, however, be applied in such thin layers as either of the varieties of silk. This often precludes its use where space economy is of paramount importance.

As a result of these qualities, cotton is often used as an outer covering of wires that have already been insulated with silk or enamel, thus combining the advantages of ease of coloring and mechanical protection with those of higher insulating value and space economy.

The most important varieties of cotton are Peeler, Sea Island and Egyptian. The first of these is the one most grown in the United States and has a relatively short fiber averaging about one inch in length. Sea Island cotton has a long fiber of superior strength and fineness which permits its use in the making of finer yarns. Its fiber lengths range from about one-half to two inches. Egyptian cotton has a fiber length between these two.

The relative advantages of silk and cotton, with respect to the insulation they afford, depend largely on the circumstances of their use, particularly with regard to the humidity to which they are exposed. Both are absorbent and, of course, the absorption of moisture will materially affect the insulating qualities of either.

The relative degrees with which silk and cotton will absorb moisture under various conditions of humidity, after having been rendered bone dry, are shown in Fig. 174 taken from the above-mentioned article. This shows, contrary to the usual belief, that silk absorbs more moisture than cotton and, therefore, might be an inferior insulator. Figure 175, however, by the same authority, shows how the power loss of a pair of cotton-insulated wires compares with that of a pair of silk-insulated wires when carrying voice currents under different conditions of atmospheric humidity. Under conditions below 70 per cent relative humidity, cotton is only slightly inferior to silk, but as the humidity increases the current leakage through the insulation increases more rapidly for cotton, than for silk and at 90 per cent humidity cotton shows a power loss about three times as great as silk.

Wool insulation behaves much like silk in this respect, while vegetable fibers, such as linen, jute or hemp, closely approximate the behavior of cotton.

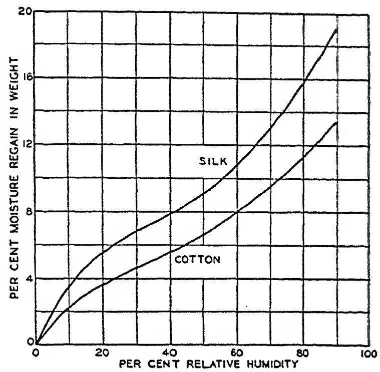


FIG. 174.—Comparative moisture regain in weight of silk and cotton insulation. (*Courtesy of Bell Telephone Laboratories.*)

The use of wool as an insulator of electrical wires is now comparatively rare. It is available, however, where high-insulating values are required without exacting requirements of space economy. Linen, jute and hemp are but little used for telephone-wire insulation.

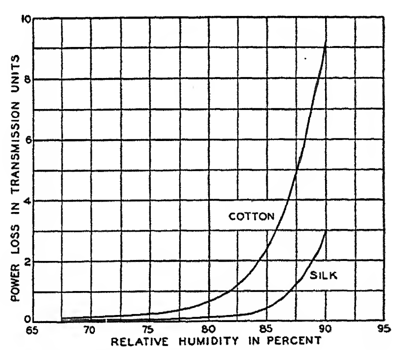


FIG. 175.—Comparative power loss in transmission units of silk and cotton insulation. (*Courtesy of Bell Telephone Laboratories.*)

Textile insulation is usually applied by closely wrapping the textile yarn about the wire. Where one wrapping does not afford sufficient insulating qualities or sufficient mechanical protection, two or more are applied, the successive layers

always being wound in opposite directions to aid in preventing unraveling. Sometimes a combination of the two materials is used, a common form being a single layer of silk for its high-insulating qualities followed by a single layer of cotton for its mechanical protection and color designation. From these various modes of applying and combining the wire coverings, we have such terms as "single silk," "double silk" and "single silk single cotton" to designate the kinds of insulation.

For some uses the textile material is braided rather than wrapped on the wire. Where this is done, it is usually in the application of an outer covering of cotton over one or more wrappings of silk. This finds its principal application in the wiring of central-office switching apparatus under conditions where many short lengths of wire are required. In this case the outer braiding has a distinct advantage over the outer wrapping, in that it is not so likely to unravel and loosen at the exposed ends. It is, however, far more costly.

Enamel, as a material for insulating wires used in connection with telephone equipment, now occupies a front-rank position. In many cases it has superseded cotton and silk in the insulation of magnet wire. It is also used in combination with coverings of silk or cotton, or both, not only in magnet wire but also in other wires formerly insulated only with textile coverings.

In the so-called "enameled wire," the bare wire is coated with a sort of varnish, compounded of gums and oils. This is applied as a liquid, but when submitted to the proper baking it becomes a hard, tough, very thin enamel, affording high insulation, great dielectric strength and excellent moisture resisting qualities.

In applying the enamel coating the wire is slowly drawn through a bath of insulating compound, heated to a temperature which gives it the proper consistency. In its continuous movement it then passes through an oven where the very thin coating attained in the bath is baked on. Thence, it again passes through the bath where it attains another coating which, in turn, is baked on when it again reaches the oven. This alternate immersion and baking is repeated as one continuous process in five or six successive stages, the wire finally emerging with the proper number of coatings successively acquired and baked.

The application of the enamel coating in successive thin films guards effectively against the occurrence of bare spots on the wire, such as "pin holes" caused by air bubbles, which might

exist in a single coating. Enamel as an insulating material is superior to either silk or cotton in two principal respects: space economy and moisture resisting qualities.

Its saving in the amount of space occupied will be set forth more specifically later on, but to illustrate at this point: Enamel insulation adds only about 0.0004 inch to the diameter of a No. 40 B. & S. gage wire, while a single layer of silk adds about 0.002 inch and of cotton 0.004 inch. The thickness of the enamel insulation, therefore, is only about one-fifth that of a single silk-covered wire and one-tenth that of a single cotton-covered wire.

In its moisture resisting properties, also, the advantages of enamel are pronounced. Coils of wire thus insulated will stand immersion in water for long periods without injury, whereas coils wound with textile insulated wire, not otherwise protected, will be ruined immediately by such immersion.

Enamel suffers somewhat in comparison with either silk or cotton in respect to its ability to stand rough handling. This is particularly true when the enamel has not been properly compounded or applied, with the result that the coating is either too hard or too soft. If too hard, it is brittle and easily flakes off. If too soft, it is unable to prevent adjacent conductors from being forced together under pressure. These dangers are not so pronounced when the enamel coating is protected by another covering of textile insulation. Wires so coated with coverings of silk and cotton over the enamel occupy an important place in telephone equipment work, much of the so-called "switchboard wire" being insulated in this way.

For many uses, however, space requirements will not permit such additional covering, and under proper conditions it is not necessary. Enamel that has been properly compounded, applied and baked is sufficiently tough and elastic to withstand any reasonable handling without injury.

Its toughness and elasticity have been developed to such a point that sizes smaller than No. 25 B. & S. gage may be stretched to the breaking point of the wire without rupturing the enamel, while the larger sizes may be stretched as much as 30 per cent without causing it to check or flake off. Sizes from No. 6 down to No. 17 may be wrapped about a mandrel of twice their own diameter, and sizes No. 18 and smaller around a mandrel of their own diameter, without flaking or cracking the enamel.

These characteristics of toughness and elasticity would be easy to attain if it were permissible to employ a relatively soft enamel. They are, however, practically attained with enamel that is so hard as to be uninjured by the pressures that are encountered in winding one layer of wire upon another.

The insulating qualities of enamel are inherently high. The real criterion of the adequacy of insulation of enamel wire, however, rests not so much on the insulation resistance and break-down strength of the enamel *per se* as on the continuity of the coating and the uniformity of its thickness. One test that has been extensively employed in an effort to detect pin holes and bare spots is the mercury bath test, the insulation between the wire and mercury being observed during immersion. This test, while apparently drastic, seems to give unconclusive and inconsistent results. Apparently due to changes in surface tension of the mercury at different voltages, low potentials develop more apparent faults in the insulation than higher ones.

Because of these seemingly inconsistent results and because the mercury bath test does not approximate conditions of use, what is known as the "layer test" may be used. This is a breakdown test between two layers of wire wound on a smooth round mandrel of hard insulating material, under sufficient tension to give a smooth winding but not enough to stretch the wire. The conditions are thus made to approximate closely those of actual use in magnet coils.

The conditions for the test, as given by the Belden Manufacturing Company, are as follows: The diameter of the mandrel is 2 inches for wires of No. 14 B. & S. gage and larger, and 1 inch for smaller sizes. In all cases the length of the winding along the mandrel is 1 inch. Alternating pressure, at commercial frequency, is applied between the two layers at an initial effective voltage of not over 120, and then gradually raised until the insulation is punctured. The voltages below are those (effective) which mechanically uninjured wire should stand under this test before puncture:

Size, B. & S. Gage	Voltage
11–25	1,000
26–30	750
31–35	500
36–37	400
38–40	300

Another breakdown test is made by uniformly twisting two short lengths of enamel wire together as shown in Fig. 176. The tension applied in twisting should be insufficient to stretch or injure the wire. For sizes No. 30 and smaller, 4-inch lengths of wire should be used, with 10 twists to the inch. Under alternating pressure, applied as just outlined, the insulation should not break down at less than the effective voltages indicated for the layer test.

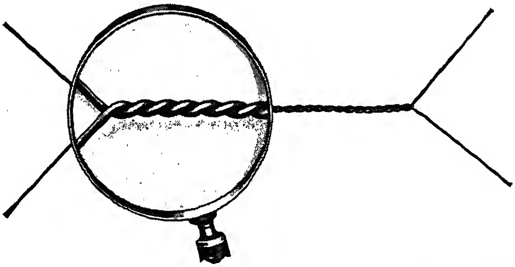


FIG. 176.—Test for enamel wire. (*Courtesy of Belden Manufacturing Co.*)

In general, the dielectric strength of enamel coating is sufficient to resist about 500 volts per mil of enamel thickness. Measured in the same way the dielectric strengths of silk and cotton coverings per mil thickness of insulation are about as follows:

	Volts
Single cotton	135
Double cotton	115
Single silk	180
Double silk	170

To recapitulate regarding enamel coating on wires: It has inherently high-insulation resistance, good dielectric strength and is impervious to moisture. It has better heat-resisting qualities than either silk or cotton and is chemically stable. It is hard enough to resist undue deformation under the pressures incurred in coil winding and is elastic and tough enough to remain intact with reasonable care in handling. It occupies less space than any of the other forms of insulation, especially for the smaller sizes of wire, and is only equalled by single silk in this respect for the larger sizes. Lastly, it is comparatively cheap.

The remaining important insulating material for interior telephone wires is rubber. By this is not meant pure rubber but some rubber compound, such as fine Para rubber mixed with such mineral ingredients as sulphur, whiting talc and litharge. Some so-called "rubber compounds" contain very little rubber. On the other hand, a compound containing no more than 30 per cent pure rubber gum compounded with other proper materials is considered a high grade of rubber insulation. In general, the insulation resistance of rubber compounds increases with the percentage of rubber content, while the mechanical qualities are improved by the other ingredients.

Rubber-covered wires for interior telephone use are usually, though not always, provided with an external covering in the form of a cotton braid. This may be variously colored for general appearance or for identification purposes. Such wires are used in switchboard work for carrying the relatively high voltage lighting or ringing currents, in house wiring on account of moisture conditions that are likely to be encountered, and in various other places where available space permits and where a high degree of moisture-resisting qualities and ability to withstand rough usage is desired. On account of its high specific inductive capacity it is not used in long lengths in voice-carrying circuits.

Having briefly discussed some of the qualities of the principal wire-insulating materials we may now turn to the types of insulated wire particularly adapted to specific uses.

Magnet Wire.—Insulated copper wire in those forms most suitable for winding electromagnet coils has come to be known as *magnet wire.* The definition cannot be made very exact because, strictly speaking, conducting wires of other metal than copper and with almost any sort of insulation or without insulation are sometimes used in the winding of such coils. According to earlier usage the term "magnet wire" was applied to the smaller sizes of round soft copper wire covered with a thin wrapping either of silk or cotton. Later, the scope of the term has been considerably broadened and now includes insulated conductors of square, rectangular or round cross-sections with coverings of enamel, silk or cotton, or combinations of these materials.

Round magnet wires are principally used in telephony, and the sizes commonly designated as magnet wire range from those having a bare diameter of about an eighth of an inch down to

those considerably smaller than a fine human hair; that is, from about No. 8 to No. 46 B. & S. gage (see Table III).

At the present time cultivated silk, cotton and enamel are practically the only materials used for insulating magnet wire. Tussah silk is but little used. Until about 1905, only silk and cotton were employed, but now enamel has gained in favor to such an extent that it is more widely used in the coils of telephone apparatus than any other insulating material. One of the principle reasons for this gain in the use of enamel insulation is the relatively small amount of space it occupies, particularly in the smaller gages of wire. The importance of space economy in coil winding will be apparent when it is considered that it is the copper conductor and not the insulation that is effective in producing magnetic results, and that the space occupied by the insulation is in that respect waste space. The thinner the insulation the smaller the space occupied by a coil of a given number of turns and size of wire. Also, the mean length of turn will be smaller. The saving, therefore, is not only in space occupied by the finished coil but in the amount of wire necessary to wind it.

Table IV shows the additions to the overall diameter of wires caused by insulating them with silk, cotton and enamel, respectively. As already pointed out, for such small magnet wires as No. 40 for instance, the thickness relationship of enamel, single-silk and single-cotton insulation is about as 1, 5 and 10, respectively. For the larger sizes, such as No. 20 and larger, enamel and silk approach equality in this respect, with single cotton considerably thicker than either. For sizes larger than about No. 16 silk is but little used.

Table V shows the approximate outside diameters of the different sizes of magnet wire after the application of the kinds of insulation indicated. These are subject to some variation, due to the difference in materials and methods of applying them employed by different manufacturers.

The last two columns of Table V concern combinations of a first coating of enamel followed by one of either silk or cotton. This additional covering is used only where space requirements will permit and where additional assurance of wire separation is required.

Magnet wires insulated with silk, cotton, enamel and various combinations of these materials each have their uses and points of superiority which may be briefly summarized as follows:

TABLE IV.—ADDITIONS TO DIAMETERS OF BARE WIRE FOR VARIOUS KINDS OF INSULATION

In inches

B. & S. gage	Single silk	Double silk
40 to 33	0.0019 to 0.0021	0.0038 to 0.0040
32 to 26	0.0016 to 0.0018	0.0032 to 0.0037
25 to 22	0.0018 to 0.0019	0.0036 to 0.0038
21 to 18	0.0019 to 0.0021	0.0039 to 0.0042
17 to 15	0.0022 to 0.0025	0.0045 to 0.0048
	Single cotton	Double cotton
40 to 33	0.0042 to 0.0044	0.0080 to 0.0083
32 to 26	0.0043 to 0.0045	0.0082 to 0.0084
25 to 22	0.0044 to 0.0046	0.0083 to 0.0085
21 to 17	0.0048 to 0.0049	0.0092 to 0.0095
16 to 11	0.0049 to 0.0051	0.0094 to 0.0096
10	0.0058 to 0.0061	0.0107 to 0.0110
9	0.0068 to 0.0070	0.0110 to 0.0116
8 to 4	0.0075 to 0.0085	0.015 to 0.016
	Enamel	
46 to 41	0.0001 to 0.0002	
40 to 36	0.0004 to 0.0005	
35 to 31	0.0006 to 0.0008	
30 to 26	0.0008 to 0.0012	
25 to 21	0.0013 to 0.0017	
20 to 16	0.0018 to 0.0020	
15 to 11	0.0020 to 0.0020	
10 to 5	0.0021 to 0.0021	

As between silk and cotton, silk is the better insulator under oridinary atmospheric conditions and may be applied in thinner layers. Cotton, on the other hand, is more hardy from the mechanical standpoint and is very much cheaper. In spite of the great disparity in costs, however, silk is often the more economical, because, being so much thinner, the required number of turns may be secured in much smaller winding space.

As between magnet wires insulated with enamel and textile materials, enamel possesses advantages with respect to space economy, moisture and heat-resisting qualities, chemical stability

TABLE V.—OUTSIDE DIAMETERS OF MAGNET WIRE

B. & S. gage	Enamel	Single cotton	Double cotton	Single silk	Double silk	Cotton enamel	Silk enamel
8	0.1306	0.1355	0.1415	......	......	0.1376	
9	0.1165	0.1214	0.1274	......	......	0.1235	
10	0.1040	0.1079	0.1129	......	......	0.1100	
11	0.0927	0.0967	0.1017	......	......	0.0987	
12	0.0828	0.0868	0.0918	......	......	0.0888	
13	0.0740	0.0780	0.0830	......	......	0.0800	
14	0.0661	0.0701	0.0751	......	......	0.0721	
15	0.0591	0.0631	0.0681	......	......	0.0651	
16	0.0528	0.0558	0.0608	0.0528	0.0546	0.0578	0.0548
17	0.0470	0.0503	0.0553	0.0473	0.0491	0.0520	0.0490
18	0.0421	0.0453	0.0503	0.0423	0.0441	0.0471	0.0441
19	0.0377	0.0409	0.0459	0.0379	0.0397	0.0427	0.0397
20	0.0337	0.0370	0.0420	0.0340	0.0358	0.0387	0.0357
21	0.0302	0.0335	0.0385	0.0305	0.0323	0.0352	0.0322
22	0.0269	0.0293	0.0333	0.0273	0.0291	0.0309	0.0289
23	0.0241	0.0266	0.0306	0.0246	0.0264	0.0281	0.0261
24	0.0215	0.0241	0.0281	0.0221	0.0239	0.0255	0.0235
25	0.0192	0.0219	0.0259	0.0199	0.0217	0.0232	0.0212
26	0.0171	0.0199	0.0239	0.0179	0.0197	0.0211	0.0191
27	0.0153	0.0182	0.0222	0.0162	0.0180	0.0193	0.0173
28	0.0136	0.0166	0.0206	0.0146	0.0164	0.0176	0.0156
29	0.0122	0.0153	0.0193	0.0133	0.0151	0.0162	0.0142
30	0.0109	0.0140	0.0180	0.0120	0.0138	0.0149	0.0129
31	0.0097	0.0129	0.0169	0.0109	0.0127	0.0137	0.0117
32	0.0087	0.01195	0.01595	0.00995	0.01175	0.0127	0.0107
33	0.0077	0.01108	0.01508	0.00908	0.01088	0.0117	0.0097
34	0.0069	0.01030	0.01430	0.00830	0.01010	0.0109	0.0089
35	0.0062	0.00961	0.01361	0.00761	0.00941	0.0102	0.0082
36	0.0055	0.00900	0.01300	0.00700	0.00880	0.0095	0.0075
37	0.0049	0.00845	0.01245	0.00645	0.00825	0.0089	0.0069
38	0.0044	0.00796	0.01196	0.00596	0.00776	0.0084	0.0064
39	0.0039	0.00753	0.01153	0.00553	0.00733	0.0079	0.0059
40	0.0035	0.00714	0.01114	0.00514	0.00694	0.0075	0.0055

and cost. It is inferior, however, in being somewhat more subject to injury by rough handling.

Where winding space permits, the application of a silk or cotton covering over one of enamel combines the insulating advantages of both these types of insulation. The resulting

superior insulation is, of course, attained at a sacrifice of space economy and cost.

Further information concerning magnet wire will be given in the next chapter on electromagnetic windings.

Annunciator Wire.—This name is applied to a soft copper wire insulated with two layers of cotton wrapped in opposite directions. The outer layer is usually distinctively colored, often red and white. After the cotton coverings are applied, they are saturated with paraffin and the whole is drawn through a smooth die to give an external polish. The usual sizes of the bare wire are Nos. 18, 20 and 22 B. & S. gage.

This wire derived its name from its early use in wiring old-fashioned annunciators. It is still widely used for door-bell wiring and similar low-voltage work. It should be used only in dry places. It finds little use in telephone work.

Office Wire.—This is a better grade than annunciator wire and is used for about the same purposes. It is variously insulated, but in a common form two wraps of cotton are applied and saturated with a so-called "weatherproof compound." These are then covered with a cotton braiding, distinctively colored, saturated with paraffin and then polished. The common sizes are 16, 18 and 20 B. & S. gage. It is little used in telephone work.

Switchboard Wire.—The wire most employed in connecting the various units of central-office equipment together is termed "switchboard wire." The bare wire is always "tinned" before the insulation is applied. To do this the bright copper wire is run longitudinally through a bath of molten tin. The tin unites with the surface of the copper, and as the wire emerges from the bath all surplus tin is wiped off before it has time to cool, leaving a continuous thin bright coating of that metal.

The purpose of this tinning operation is to facilitate soldering the wire in the making of joints. No matter where the insulated wire is cut and skinned for the purpose of connecting it to an apparatus terminal or to another wire, it always presents an excellent soldering surface. The importance of this will be more apparent when the subject of joints is considered in another chapter.

The insulation of switchboard wire always consists of one or two wrappings of tussah silk and an outer covering of cotton, the whole impregnated with hot paraffin or beeswax. In some cases,

as an extra precaution, a coating of enamel is applied to the tinned wire before the textile coverings are put on.

Usually the outer cotton covering is a wrapping, but, in some cases, where the wire is used in very short lengths, it is put on in the form of braiding. This obviates the trouble which sometimes occurs with outer wrapped coverings, due to the tendency of the insulation to unwrap at endings where the wire is skinned, thus unduly exposing the wire and causing an untidy appearance. Switchboard wire in which the outer covering is a braiding of cotton is often called "bank wire" because it is largely used in the numerous short connections between the multiple-bank contacts in automatic switchboards.

Number 22 B. & S. gage is by far the most common size. Other sizes, however, are sometimes used; No. 24 where there is great need of saving in space, and Nos. 18, 19 and 20 when greater carrying capacity or a more rugged wire is required. The following are the approximate lengths in feet per pound for these sizes, with the usual types of insulation:

Switchboard Wire

Feet per pound, single conductor

B. & S. gage	1 silk, 1 cotton	2 silk, 1 cotton
18	170	160
19	200	190
20	250	230
22	360	340
24	530	480

Of recent years much switchboard wire, after being tinned is given a coating of enamel before applying the textile covering. This is termed "enameled switchboard wire." It affords a very high grade of insulation, particularly on account of its moisture-resisting qualities. It carries with it, however, one inherent objection in that it increases the difficulty of properly soldering the joints where it is connected with the apparatus terminals. Opinion at this time differs as to whether the gain in insulation qualities afforded by this practice is, under ordinary circumstances, worth while in view of the increased likelihood of the occurrence of improperly soldered joints.

Switchboard wire is ordinarily furnished either as a single conductor or with two or three insulated wires twisted together with about a 3-inch lay. These types are commonly referred

to as "single conductor," "twisted pair" or "duplex" and "triple conductor" or "triplex." The outer coverings of the individual wires are variously colored to facilitate identification, a most important feature in switchboard work. Two pieces of twisted pair, or duplex, switchboard wire are shown in Fig. 177, the one above having a braided, and the one below a wrapped, outer covering.

Switchboard Cable.—Where a sufficient number of switchboard wires are to follow the same route for considerable distances from one part of the central-office equipment to another, they are formed into cables in various combinations of single conductors,

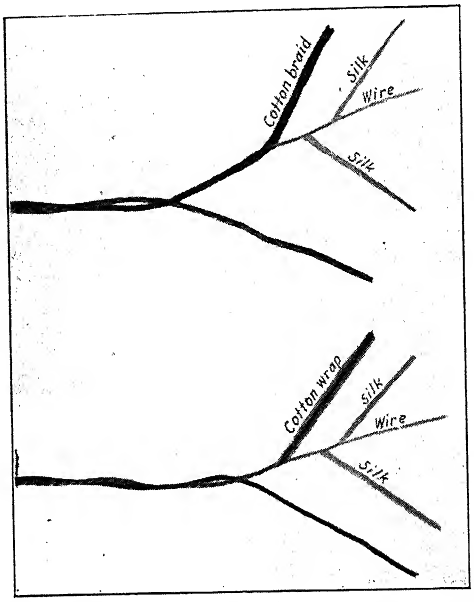


FIG. 177.—Duplex switchboard wire. Above, braided. Below, wrapped (*Courtesy of Stromberg-Carlson Telephone Manufacturing Co.*)

twisted pairs and triplets. The size of conductors is in most cases No. 22 B. & S., with some use of No. 24 where space conditions are particularly exacting. Occasionally, a few larger wires such as Nos. 18, 19 or 20 will be incorporated with the smaller wires of a cable where additional current-carrying capacity is required of some of the conductors.

The wires of switchboard cables should always be tinned. As in the case of switchboard wires that are not cabled, they are

insulated with one serving of silk and one of cotton, with two servings of silk and one of cotton or with enamel and one serving of silk and one of cotton. Braided outer covering for the individual wires is not used, since with factory made cables the condition of very short wire lengths does not often occur.

The color coded wires, after being saturated with beeswax, are laid up into a core containing the required number of conductors in the proper combinations. This core is served with a spiral of cotton thread merely to bind it together and is then closely wrapped with one or two layers of Manila paper. It is then ready for the outer covering.

Two general methods of outer covering are employed. One, largely used by the independent telephone manufacturers, consists merely in covering the Manila paper wrapping with a braiding of cotton, saturated either in beeswax or a gray fire-resisting paint. In the other, used by the Western Electric Company, a close wrapping of lead foil is placed over the core, after which the cotton braid is applied and saturated with the gray fire-resisting paint. This latter form of covering is the standard of switchboard cable generally employed by the Associated Bell companies.

Practice differs among the independent manufacturers as to the finishing of the outer cotton braiding. Some make it of a combination of red-and-white cotton yarn after which the cable is run through a bath of hot beeswax and then drawn through a polishing die to impart the required finish. The resulting red-and-white covering has led to this being called "peppermint candy cable" it being suggestive of old-fashioned stick candy. The gain in first appearance resulting from this treatment is questionable. If gain there is, it disappears with age, as the covering becomes disfigured by handling, dirt and dust. The better outer finish is the white-cotton braiding afterward saturated with dull-gray fire-resisting paint. This in the long run affords a neater appearance and is less inflammable than the highly colored covering treated with beeswax.

In case of severe moisture conditions an outer lead sheath is sometimes applied instead of, or in addition to, the outer braid. This is not, however, ordinarily necessary under central-office conditions.

The question of color code, important in single, duplex and triplex switchboard wire, is even more so in switchboard cables. The standard colors employed are blue, orange, green, brown, slate, white, red and black, either in solid colors or in various combinations.

On account of the fact that switchboard jacks in a manual multiple switchboard are commonly mounted in groups of twenty, the practice has grown up of basing the color code in a cable on a sequence of twenty readily distinguishable colors or color combinations. Such a color code for a unit of 20 pairs is as follows:

Pair Number	Designating Color
1	Blue
2	Orange
3	Green
4	Brown
5	Slate
6	Blue-white
7	Blue-orange
8	Blue-green
9	Blue-brown
10	Blue-slate
11	Orange-white
12	Orange-green
13	Orange-brown
14	Orange-slate
15	Green-white
16	Green-brown
17	Green-slate
18	Brown-white
19	Brown-slate
20	Slate-white

Only one wire of a pair carries the designating color, which distinguishes that pair from the others in a unit of 20 pairs. In larger cables this code is repeated for each unit of 20 pairs, similarly colored wires in different units of twenty being distinguished from each other by coloring the "mate" wires differently for each of the several twenties. It will be noticed that the colors red, black and solid white are not used in the 20-pair code. These colors are reserved for the mate wires of twisted pairs for distinguishing between groups of twenty in the same cable. Thus in a cable of only 20 pairs, each of the designating wires would carry a white mate. In a 40-pair

cable, pairs numbered 1 to 20 would carry a white mate and pairs numbered 21 to 40 a red mate. For larger cables up to one hundred, the designating wires of each twenty would be according to the code, the third twenty, numbers 41 to 60,

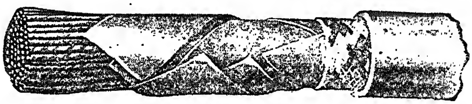


FIG. 178.—Switchboard cable. (*Courtesy of Kellogg Switchboard and Supply Co.*)

would carry a black mate, the fourth twenty, 61 to 80, a red-white mate and the fifth twenty, 81 to 100, a black-white mate.

It is customary to embody in each cable a few wires in addition to the actual requirements. These are known as "spares" and are for use in case one of the regular wires or pairs becomes

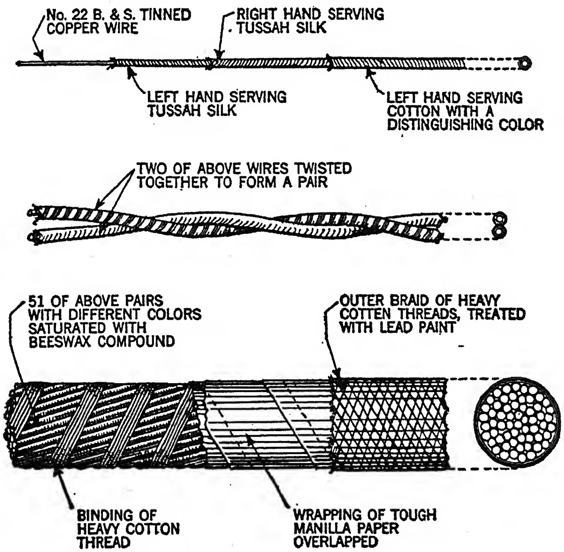


FIG. 179.—Switchboard cable. (*Courtesy of Stromberg-Carlson Telephone Manufacturing Co.*)

defective. Thus it has come about that a cable intended primarily for 10 pairs will be known as an 11-pair cable, one for 20 pairs, as a 21-pair cable, one carrying 20 pairs and 20 singles would be a 63-wire cable, and so on; this accounts for the frequent occurrence of odd numbers in referring to the usual sizes of switchboard cable. The wires of spare pairs are usually designated by a solid red with white mate for the first pair and a solid

black with white mate for the second spare. Individual spare wires are colored red-white for the first spare and black-white for the second.

A piece of 42-pair switchboard cable with coded conductors, made by the Kellogg Switchboard and Supply Company, is shown in Fig. 178. In this particular piece the red-and-white cotton braiding, which usually constitutes the outer covering, is supplemented by an outer sheath of lead. Figure 179 shows in greater detail how the 51-pair switchboard cable of the Stromberg-Carlson Company is made.

The weights per 1,000 feet for several sizes of switchboard cable variously insulated and encased are about as follows:

SWITCHBOARD CABLE, NO. 22 B. & S., GAGE

Pounds per 1,000 feet

Pairs	Braided, paraffined or flameproofed			Lead covered, no braiding
	Enamel, 1 silk, 1 cotton	1 silk, 1 cotton	2 silk, 1 cotton	2 silk, 1 cotton
11	67	66	70	319
16	90	89	94	377
21	120	118	122	450
26	150	148	154	517
42	250	245	260	832
51	280	275	310	930
102	530	520	540	1,385

Jumper Wire.—Jumper wire is so called because it is particularly adapted for use in running the jumpers or cross-connections on distributing frames. On the main frame, for instance, they bridge the gap between the outside lines terminating on one side of the frame and the "inside" or switchboard lines terminating on the other side. These jumper connections are necessarily subject to comparatively rough handling. They are of a semipermanent nature, and this leads to a considerable amount of wear and tear caused by the withdrawal of old wires or the running of new ones. For this reason, a somewhat more rugged wire than switchboard wire is desirable. Furthermore, the nature of their use on the distributing frame requires that they lie in loose masses rather than being packed closely together as in switchboard cables. They therefore present a

fire risk greater than that of most other types of wiring. This calls for fire-resisting qualities, at least to the extent that applied heat will not cause them to blaze.

The conductors of jumper wires are usually of No. 22 and sometimes of No. 20 B. & S. gage solid wire, tinned to facilitate soldering.

Sometimes wire having a single-silk, single-cotton insulation with an outer paraffin-soaked cotton braid is sold as jumper wire. This is a poor grade and is to be avoided, if for no other reason, because of its inflammability. A jumper wire of high quality made by the Stromberg-Carson Co. is shown dissected in Fig. 180. Here the outer covering is a heavy braid of cotton treated with a solution which renders it "flame-proof."

In some cases a coating of enamel is given the tinned wire before the application of the textile coverings. Under ordinary climatic conditions there is question as to whether the enamel covering is warranted. While it undoubtedly exhances the insulating qualities of the covering, it also adds to the difficulty of soldering.

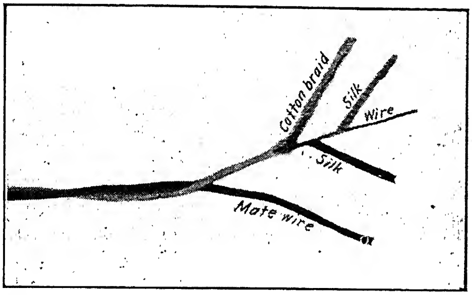


FIG. 180.—Duplex flame-proof jumper wire. (*Courtesy of Stromberg-Carlson Telephone Manufacturing Co.*)

Another good but somewhat more expensive jumper wire has an inner insulation of high-grade rubber compound covered with a tight cotton braid treated to make it fire resisting. This has been largely used. Obviously the rubber covering should be of excellent quality to prevent its becoming brittle with age.

Twisted pair and triplex jumper wires are most common, but single conductor and quadruplex are also used.

Subscribers' Station Wire.—Under this heading may be grouped wires that are commonly referred to by such trade names as "inside wire" and "interior wire." They have been

developed for such purposes as wiring subscribers' premises, that is, wiring from the point where the outside line enters the building to the telephone set and to the ground connection, if ground is required.

Subscribers' premises include not only residences but stores, factories, hotels and various other institutions requiring telephone service. Their wiring, therefore, is subject to many vicissitudes, among them rough usage, as from the moving of furniture, and moisture from such things as leaky pipes and radiators, open windows and housemaids' mops. For such reasons subscribers' station wire is usually made heavier than switchboard wire, No. 19 B. & S. gage being common. It is insulated with a good rubber compound and covered on the outside with a polished cotton braid which is left "dry," that is, not impregnated.

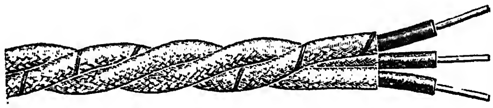


Fig. 181.—Triple conductor substation wire. (*Courtesy of Stromberg-Carlson Telephone Manufacturing Co.*)

As substation wiring is usually not concealed, and as it is frequently installed in fine offices and residences, its appearance is an important factor. An olive color is perhaps most common since it is unobtrusive and blends with almost any color of interior finish. Other colors are sometimes used in order to better harmonize with fine interior wood work. Oak and maroon are useful substitutes for olive.

Single-conductor, twisted-pair and triple-conductor interior wires are commonly supplied to the wireman. The braided covering of each wire is provided with a colored thread to act as a tracer. The tracer colors often used are: one conductor, yellow tracer; two conductors, red and green tracers; three conductors, red, green and yellow tracers.

A triple conductor substation wire is shown in Fig. 181. Such wires weigh about 11 pounds per conductor per 1,000 feet.

CHAPTER XII

COILS AND ELECTROMAGNETS

Without the electromagnetic coil or helix, electric telephony in anything like its present state could not exist. It is the device which is used to translate the energy of an electric current into the mechanical pull of magnetic attraction, as in the electromagnet and in the telephone receiver; and, conversely, it is the device which translates the energy of mechanical motion into that of electrical currents, as in the dynamo and in the magneto-transmitter. But it also has functions where no mechanical pulls or motions are required. Owing to the property of self-induction, the electromagnetic helix serves to retard the sudden stopping or starting of electric current, as in the impedance coil. Also, by mutual induction, fluctuating currents in one coil cause other currents to flow in another associated coil, as in that family of plural-helix devices variously known as "induction coils," "repeating coils" and "transformers."

In all these cases the coil or helix of wire acts as the link between electric and magnetic action. They may all be included under the broad caption of "electromagnetic coils." As they form the basis of all electromagnetic devices, a brief discussion of the methods used in their construction is first in order, leaving for later consideration the kinds of cores upon which they are wound and the kinds of functions they are to perform.

The structure upon which the coil is to be wound may be in the form of a spool or bobbin, in which case the inner surfaces of the spool heads definitely limit the winding space at its ends. For other types of winding the structure becomes merely a tube of insulating material, in which case the formation of the winding itself determines the end limitations.

With the spool or bobbin type of winding, the shank of the spool may be formed by the iron core that is to be magnetized by the coil, or it may be formed of a tube of insulating material into which the core is slipped after the coil is wound. In either case, the spool heads, usually of red fiber or phenol fiber,

are slipped on to the ends of the shank. They are merely forced on to the roughened surfaces of the iron cores as in Fig. 182, or glued on to the ends of the insulating tube as in Fig. 183.

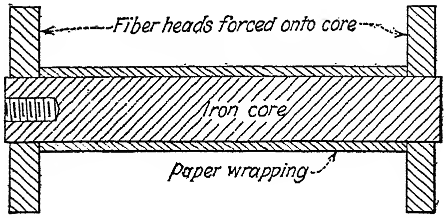


FIG. 182.—Bobbin formed on core.

In all cases where, as in Fig. 182, the iron core forms the shank of the spool, it is necessary to surround it by one or more layers of insulating material, usually paper, in order that the winding

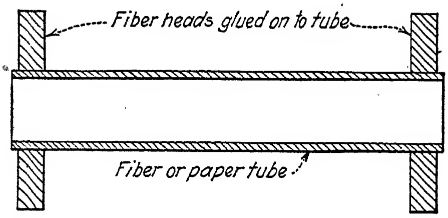


FIG. 183.—Bobbin formed on insulating tube.

may be more effectively insulated from the core than if it were wound directly on the core without this protection.

Since the advent of the various phenol fiber compounds, which with heat and pressure may be molded into intricate forms, molded spools or bobbins have found increasing use. One of these is shown in Fig. 184. In the particular case illustrated, the iron core is placed as an insert into the mold, and the phenol fiber is molded around it. The core thus forms an integral part of the bobbin. The bobbin may, with equal facility, be made of tubular form, for the later insertion of the core. This molded type of spool construction is somewhat more expensive than the older types, but it has advantages which warrant its use under some conditions. Of course, such a spool possesses very high-insulating

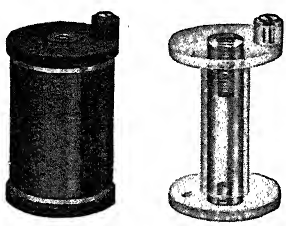


FIG. 184.—Molded bobbin.

qualities. An additional advantage is the great accuracy with which such molded parts may be made. The smoothness and freedom from projections which may be given to the inner faces of the spool minimize danger of abrading the wire during the winding process.

With bobbins or spools two general forms of winding are used. These, according to their method of application, may be referred to as "hand" and "machine" windings. The method termed "hand winding" is not to be taken as meaning that the wire is wrapped on by hand without the aid of machinery. It means,

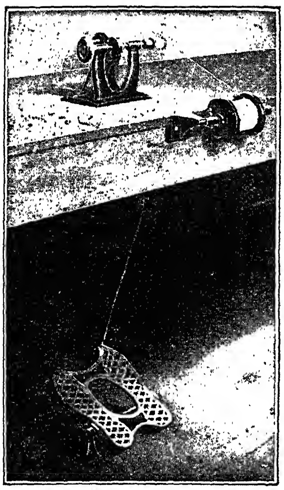


Fig. 185.—Winding machine.

Fig. 186.—Revolution counter for coil winding.

rather, that the wire is guided on by hand while the spool is being rapidly revolved by a so-called "winding machine." These machines may be driven by foot power, as with an ordinary sewing machine treadle, but where quantity production is involved, they are always power driven.

A common form of hand-winding machine is shown in Fig. 185, this being one of the several types made by the Belden Manufacturing Company. This machine derives its power by belt from a line shaft shown under the bench which may be common to many such machines. The belt drives a short shaft mounted on a rocker arm parallel to and just behind the winding spindle. This short shaft carries a friction driving pulley adapted

to engage a similar driven pulley on the winding spindle when the rocker arm is in its forward position. When in its rearward position the rocker arm applies a brake to the driven pulley. The position of the foot treadle controls that of the rocker arm. An intermediate position of the rocker arm and of the foot treadle disengages both pulley and brake and leaves the spindle free to revolve by hand in either direction, which is particularly useful in starting and finishing the winding and also in backing up to correct an error.

As will be shown, the number of turns on a winding is an important factor in determining the electromagnetic performance. A useful adjunct to winding machines, therefore, is a revolution counter, one of which is shown in Fig. 186. This is usually connected to the winding spindle by a flexible shaft and thus indicates the number of turns by the two pointers which move around the dial. These pointers are friction mounted so that they may be set to zero at the start of each winding.

If in the winding operation little regard is paid to the securing of uniform layers, the process is referred to as "random winding." Usually, the better practice in hand winding is to wrap the wire as nearly as possible in even layers. Complete success in this operation is difficult, particularly with the finer wires. It has become common practice, therefore, after winding each few layers, to introduce a layer of paper to form a new surface for the succeeding layers.

The reason even winding in layers rather than random winding is desirable is threefold: First, winding in layers is more compact, permitting more turns in a given space and more turns for a given length of wire, thus promoting both space economy and wire economy. Second, it is obvious that greater differences of potential will ordinarily exist between the turns of adjacent layers than between adjacent turns of the same layer. Also, greater differences of potential will exist between turns in layers that are not adjacent than between those in adjacent layers. If, therefore, due to irregularities, one of the turns that ought to be in a layer near the outside of the coil crowds down through the turns of preceding layers so as to lie near a turn in one of the inner layers, it constitutes a danger point due to the relatively large differences of the potential that may exist between the turns thus improperly brought into close proximity. Third, where the windings are evenly laid on in layers, the pressure

between the turns of successive layers is more uniformly distributed than if the wires are allowed to crisscross. Where one wire crosses over another, the pressure between the two is concentrated at a point and is likely so to crush the insulation as to bring the wires into dangerous proximity if not into actual contact.

In the machine method of winding the bobbin is held on a rapidly revolving spindle, the wire being fed over a small sheave that is automatically moved back and forth along the length of the winding space. The arrangement in this respect is similar to that of the automatic feed on a lathe, except that in coil winding the direction of feed is automatically reversed at the end of each layer.

The amount of feed for each revolution of the bobbin is accurately adjusted to the diameter of the wire, so that uniform layers are more closely approximated than by hand winding and greater speed is attained. The automatic or machine winding of bobbin coils, however, is fraught with considerable difficulty. The rigid mechanical limitations imposed by the bobbin leave no room for adjustment to provide for such variations in the outside diameter of magnet wire as are inevitable in its manufacture. While machine winding on bobbins is successfully used where very large production is involved, these difficulties have worked against its general adoption, with the result that hand winding is still used to a very large extent on bobbin coils.

A form of winding machine known as the "semiautomatic," which successfully meets the requirements of machine winding on bobbins, has been devised. In this the wire is automatically fed onto the spool. The reversal of the feed movement is brought about automatically when the inner cheek of the bobbin is reached at the end of each layer. In this way the time of feed reverse is independent of such variation in the number of turns as may be brought about by variations in the diameter of the wire or by slight variations in the distance between spool heads.

With the semiautomatic machine, the winding goes on automatically until some irregularity develops, at which time the operator introduces a layer of paper in order to obtain a fresh start for the next layer. These machines, while requiring the undivided attention of an operator, permit high speeds and give excellent results. The facts, however, that they require the undivided attention of an operator and involve a very much

greater investment make it possible for the hand-winding machines to compete successfully with them in point of cost, except in cases where large production is involved.

As distinguished from coils wound on bobbins or spools, there is to be considered the other important class in which the coil is wound on an insulating tube, without any structural barriers, like spool heads, to establish the end planes of the winding. These, as now employed in telephone practice, are mainly of two classes: paper-section coils and universal coils. These may be called "tube windings" because they are wound on insulating tubes without spool ends. The tubes may or may not contain iron cores at the time of winding.

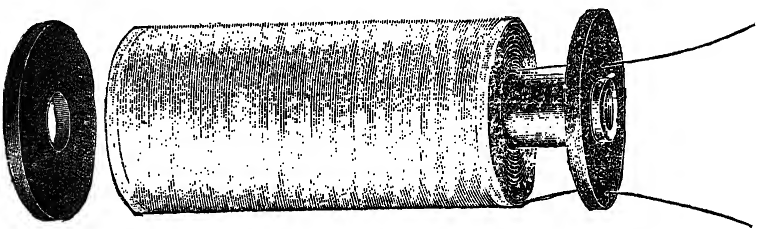


Fig. 187.—Paper section coil.

These tube windings are practically always machine wound. There being no definite limitations imposed by spool ends, the machine feed may be reversed at the end of each definite number of turns, the slight variation of magnet wire diameter merely resulting in a slight variation in the length of the layers The tube is usually composed of one or more thicknesses of strong pressed paper giving a structure of considerable rigidity.

In the so-called "paper-section coil," a layer of thin insulating paper is fed into the coil automatically at the end of each layer. The width of these paper inserts is such as to project about $\frac{1}{16}$ to $\frac{1}{8}$ inch beyond each end of the actual winding, so as to give firmness and prevent the layers from breaking down at the ends.

The machines for making these coils are entirely automatic and work with astonishing rapidity and accuracy. In many cases a single machine will wind a number of coils at once, these being spaced uniformly along a long paper tube placed on a single mandrel. After winding, the paper tube is cut into sections, thus separating the individual coils. Such a winding is shown in Fig. 187 in the process of being slipped upon its core. The

spool heads shown in this figure are for protective purposes only and are sometime dispensed with.

The other method of tube winding derives its name "universal" from the machine originally developed for winding yarn into cops in the textile arts. These so-called "universal machines," as used in winding cotton yarn, for instance, feed the yarn along the length of the tube by a very rapid transverse motion, so that the yarn crosses itself every few turns. As a result of this diagonally crisscrossed disposal of the yarn, the cop so formed is self-supporting at its ends.

In the application of this universal machine to coil winding processes, a separate feed for the magnet wire has been added to the feed for the yarn. This magnet wire feed lays on the wire in close uniform layers with a spacing determined by the wire diameter. Meanwhile, the yarn feed operates in the usual way to lay one or more strands of fine cotton yarn diagonally from one end of the winding to the other. This binds the successive layers of wire together and builds up a winding that is even throughout its length and firm at the ends. Such a winding is shown in Fig. 188.

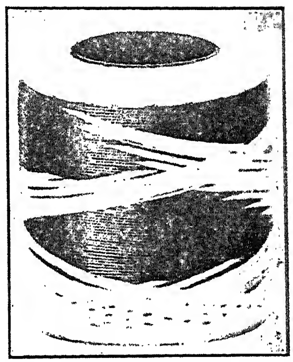


Fig. 188.—Universal winding.

Enamel wire is nearly always used for this purpose and although some space is taken up by the cotton yarn, the space factor is generally better than that realized in ordinary windings of silk- or cotton-insulated wire. The diagonally laid threads of cotton interposed serve to bind the winding together longitudinally and, also, to facilitate impregnation, if that is to be done later.

The cotton interwinding projects from $\frac{1}{16}$ to $\frac{3}{16}$ inch beyond each edge of the winding, thus preventing the breaking down of the end turns and affording protection of the wires constituting the layer ends. To afford protection of the outer layer, the cotton winding may be continued after the last layer of wire is reached, until a sufficient covering is provided. Universal windings may be wound directly on the iron core, if preferred, but in this case the core must, of course, be insulated with a

wrapping of paper to prevent short circuits between it and the inner turns of the winding.

It is possible to wind magnet coils with bare wire. Such a method was invented by Dr. Leverett Bradley, in 1865, and about the beginning of the present century was developed to a high state of efficiency by the Varley Duplex Magnet Company. The method is indicated in Fig. 189 and consisted in winding a thread of silk or cotton alongside and parallel with the wire, so as to maintain the required lateral separation. The separation between layers was provided by a thickness of paper introduced between the successive layers. In developing this type of coil and the technique of its making, the Varley Company laid the foundation for the present sectional winding coil. The automatic machines perfected by that company form the basis of much of the modern magnet wire machinery of today.

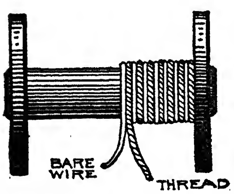


FIG. 189.—Bare wire winding.

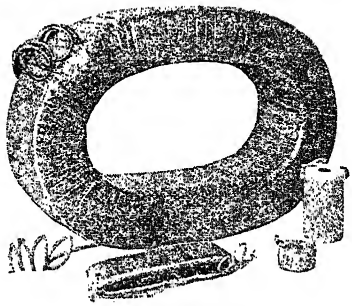


FIG. 190.—Form-wound coil. (*Courtesy of Belden Manufacturing Co.*)

Another type of electromagnet coil winding is found in the form-wound coil. These are coils that are wound on forms or spools which are afterward removed. Such a coil is shown in Fig. 190. These are much used where irregular shapes are required and where the wire is large enough to afford sufficient stiffness to the coil to prevent its distortion when left unsupported. After removal from the forms they are taped for protection and outer insulation, after which they are applied to the core structure with which they are to be associated.

Due to comparatively recent developments in magnetic core material, it is probable that the form-wound coil will play a part of increasing importance in telephony in the future. The so-called "dust core" is formed of iron so finely comminuted as to permit its being molded under pressure into required

shapes. The thought of forming the core around the coil instead of the coil around the core presents alluring possibilities.

An important phase of the technique of coil construction is the bringing out of the terminals to provide external connection with the ends of the magnet wire. Most of the coils used in telephony are wound with comparatively fine wire. This is fragile and likely to break off if left projecting from the coil without support or reenforcement. The breaking off of a terminal close to the body of the coil, particularly if it be the terminal of the inner end, usually means the loss of the coil.

To guard against this, the ends of the magnet wire are reenforced by short pieces of heavier wire or, better, by a very flexible conductor known as "lead-in wire" or "lead wire." This

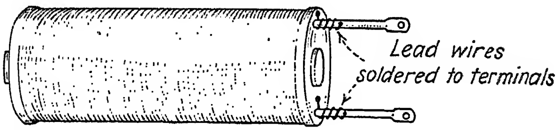


Fig. 191.—Complete electromagnet with terminals.

consists of a number of strands of fine copper wire laid up by twisting or braiding into a small cable. The individual wires are usually tinned to facilitate soldering. The resulting conductor is insulated with a double layer of silk or of silk and cotton.

Before starting the winding the end of the magnet wire is soldered to the end of a lead-in wire and, after covering the joint, usually with a fold of paper for insulation, the winding is started in such manner that its first few turns will be formed of the lead wire. In this way the inner end of the fine wire forming the coil is anchored and protected, the only exposed wire being the flexible lead wire. At the completion of the winding another lead wire is similarly attached before the outer covering of the coil is applied.

The lead wires projecting from the body of the coil may be formed into "pigtails," in case flexible terminals are desired. More often, however, they are directly soldered to rigid terminals, either mounted directly on the spool heads or on the base of the apparatus, if a base is provided. A common form of coil with terminals on the spool head is shown in Fig. 191.

Here the lead wires are brought out through the spool head, wrapped around the parts, and soldered to them.

The question of soldering in such places as these will be referred to at greater length in Chap. XVI on Joints and Contacts. The subject is so important, however, that it is well to anticipate, at this point, by saying that the greatest care should be taken to avoid the use of any soldering flux that would tend to eat away the wire by corrosion or to lower the insulation by gradually creeping into the insulating material. Either of these occurrences may be disastrous, particularly in telephone switchboards where very large numbers of fine wire coils are grouped together in comparatively small space. On these accounts, it can not be too strongly impressed that the soldering flux employed generally in connection with telephone apparatus should consist only of rosin. This is not only non-corrosive but also is, of itself, a good insulator.

Opinion differs as to the desirability of impregnating telephone coils with some sort of insulating material which will fill all interstices and become hard after cooling. Impregnation is most often practiced with paper-section or universal coils wound with enamel wire. It is, of course, of the utmost importance that, if coils are to be impregnated at all, no ingredients or heating methods shall be used which will dissolve or otherwise injure the enamel coating on the wire. The facts that impregnation compounds are usually some form of varnish and that the enamel coating is essentially a dried varnish make it necessary to consider carefully the kind of solvent used in the impregnating fluid. A varnish thinned with benzine is the safest, because benzine is the mildest available solvent. Coal-tar solvents and the alcohols are stronger and should not be used, because of danger that they may attack the enamel coating.

While there is undoubtedly great benefit to be derived from coil impregnation in other arts, it has been the writer's experience that, in telephone work, impregnation causes quite as many troubles as it prevents, and there is little need for it. Electromagnet coils as used in switchboards are usually grouped closely together in large numbers and are in close proximity to the delicate contacts which have to be made and broken in the normal operation of the switchboard. Sometimes a winding becomes hot, either due to the flowing of a normal current for an abnormally long time or to the flow of an abnormal current even

for a short time. In some cases the emanations from an impregnated coil under such conditions have caused much trouble. This trouble is particularly elusive where the emanations occur in the form of a volatile gas, which, in time, may condense on the contact points in sufficient degree to interfere with their proper operation.

It is understood that the Western Electric Company, by far the largest manufacturer of such coils, has now adopted the policy of impregnating none of its telephone relay coils. This applies even to those that are to be sent into tropical climates where great humidity often prevails.

The most important factors in the design of an electromagnet coil for a given purpose are the dimensions of the winding space, the resistance of the winding and its number of turns. The dimensions, of course, determine the volume of the winding space. This, in turn, subject to variations in insulation thickness and method of winding, determines the number of turns of a given size and kind of wire that may be secured. On the number of turns depends the length of wire and its resistance.

The resistance of the winding determine the amount of direct current that will flow through it for a given electromotive force. On the number of turns depends the amount of magnetizing force that a given current will develop.

The coil dimensions are largely influenced by considerations of mechanical design, always, of course, within the limitations imposed by the feasibility of doing what is wanted within the given space. While it is possible to predetermine the actual performance of a coil, as, for instance, the pull of an electromagnet, this is so largely affected by the mechanical characteristics of the piece of apparatus in which the coil is to be used and on the magnetic characteristics of the core and armature structure that the performance capabilities are almost always determined experimentally.

The general dimensions of the coil having been tentatively determined, questions arise as to how the required number of ampere turns may be most effectively and economically provided. The factors of space, resistance and number of turns are closely interrelated. Coil calculations may be conveniently worked out in connection with such a diagram as that of Fig. 192, in which the various pertinent dimensions are given the designating letters that have, by general adoption, become almost standardized for

such work. These dimensions all refer to the *winding space* afforded by a cylindrical spool.

Obviously, the number of turns that may be secured in a coil of given dimensions depends not only on the gage of wire but also on the character of its insulation. It is the outside diameter of the insulated wire that determines the space it occupies. Table V, already referred to in Chap. XI, gives the outside diameters of magnet wires of the B. & S. gage numbers most used in telephony and for the various kinds of insulation. Too much accuracy must not be expected of such a table, as, obviously, variations will occur, particularly as between the wires of different manufacturers. This and the following tables of this chapter were furnished by the Belden Manufacturing Company.

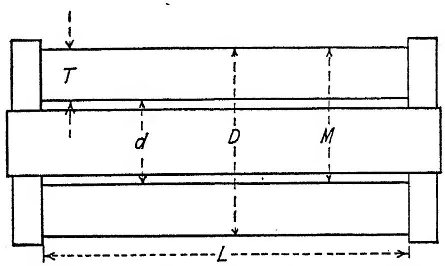


Fig. 192.—Winding-space dimensions for cylindrical coil.

From the wire diameters of Table V the number of turns that can be placed in a layer is easily determined from the length L of the winding space, and the number of layers from the depth or thickness T. The product of the turns in a layer and the number of layers will, of course, be the number of turns in a coil. Thus, the question of how many turns of a given size of wire may be placed in the cross-sectional area $T \times L$ is merely a matter of arithmetic when the diameters of Table V are used.

Much of this arithmetical work has already been done, and the results collected in Table VI. This shows the number of turns that can be put in a square inch of winding cross-section for each size and kind of magnet wire. The figures of Table VI are in each case something less than the reciprocals of the squares of the corresponding diameters of Table V, due to the practical conditions met in winding, such as irregularities and paper inserts. For given winding space dimensions, it is only neces-

TABLE VI.—TURNS PER SQUARE INCH

Symbol n

B. & S. gage	Enamel	Single cotton	Double cotton	Single silk	Double silk	Cotton enamel	Silk enamel
8	57	53	48	......	......	52	
9	72	66	59	......	......	64	
10	90	84	76	......	......	80	
11	113	104	93	......	......	100	
12	141	129	114	......	......	124	
13	177	160	140	......	......	151	
14	221	198	171	......	......	187	
15	277	245	208	......	......	230	
16	348	312	260	351	327	289	326
17	437	383	316	437	405	358	408
18	548	472	378	548	503	438	505
19	681	581	455	682	619	532	622
20	852	712	545	848	761	644	769
21	1,065	868	650	1,055	935	780	946
22	1,340	1,128	865	1,315	1,150	1,008	1,175
23	1,665	1,370	1,030	1,620	1,400	1,220	1,440
24	2,100	1,665	1,215	2,010	1,705	1,475	1,775
25	2,630	2,020	1,420	2,470	2,070	1,790	2,180
26	3,320	2,445	1,690	3,005	2,510	2,155	2,680
27	4,145	2,925	1,945	3,680	3,010	2,590	3,275
28	5,250	3,500	2,250	4,600	3,620	3,100	4,030
29	6,510	4,120	2,560	5,530	4,270	3,660	4,865
30	8,175	4,900	2,930	6,810	5,100	4,320	5,890
31	10,200	5,770	3,330	8,260	6,010	5,120	7,170
32	12,650	6,700	3,720	9,870	6,990	5,960	8,560
33	16,200	7,780	4,140	11,850	8,160	7,020	10,400
34	19,950	9,010	4,595	14,250	9,480	8,060	12,200
35	25,000	10,300	5,070	16,800	10,870	9,200	14,500
36	31,700	11,750	5,550	19,850	12,430	10,550	17,300
37	39,600	13,250	6,045	23,300	14,100	12,000	20,400
38	49,100	14,900	6,510	27,300	15,960	13,400	23,600
39	62,600	16,600	6,935	31,700	17,850	15,150	27,850
40	77,600	18,400	7,450	36,700	19,900	16,750	32,000

sary, in order to find the number of turns on a coil, to determine the winding cross-section, TL, and multiply this area by the the number of turns per square inch, shown in Table VI for the kind and size of magnet wire chosen.

Knowing the number of turns in a coil, the total length of wire may easily be found by multiplying the mean turn length πM by the number of turns. The coil resistance may then be found from resistance tables, such as Table III. As the lengths resulting from the foregoing calculations are usually expressed in inches, however, it is convenient to know the resistances of the various gages of wire in this unit. This information is given in Table VII, in which the values correspond exactly to those of Table III.

The foregoing methods of computation are based on the number of wires that can be placed in the cross-section of the winding space. Another way of arriving at the same facts is by volumetric considerations. The volume of the space occupied by the winding may be found by ordinary geometric considerations, one of which is to multiply the length of the mean turn πM by the area of the winding cross-section TL.

Knowing the volume in cubic inches, it remains to learn how much resistance of a given size and character of wire can be placed in a cubic inch of space. In other words, we need to know the total resistance of all the pieces of wire, each an inch long, that could be laid up in a bundle of one square inch cross-section. This is readily computed for any size and kind of wire by multiplying the resistance per inch of wire (Table VII) by the turns per square inch (Table VI). For convenience, the arithmetic of this has been worked out and is given in Table VIII.

TABLE VII.—RESISTANCE PER INCH OF WIRE

Symbol s

B. & S.	Ohms	B. & S.	Ohms	B. & S.	Ohms
8	0.0000552	19	0.0006698	30	0.008583
9	0.0000659	20	0.0008450	31	0.01082
10	0.0000831	21	0.001065	32	0.01365
11	0.0001047	22	0.001343	33	0.01722
12	0.0001322	23	0.001693	34	0.02171
13	0.0001666	24	0.002136	35	0.02736
14	0.0002101	25	0.002692	36	0.03452
15	0.0002649	26	0.003396	37	0.04352
16	0.0003341	27	0.004281	38	0.05487
17	0.0004212	28	0.005399	39	0.06920
18	0.0005312	29	0.006809	40	0.08725

TABLE VIII.—RESISTANCE PER CUBIC INCH

Symbol r

B. & S. gage	Enamel	Single cotton	Double cotton	Single silk	Double silk	Cotton enamel	Silk enamel
8	0.00315	0.00293	0.00265	....	....	0.00287	
9	0.00475	0.00435	0.00388	....	....	0.00422	
10	0.00748	0.00698	0.00631	....	....	0.00664	
11	0.01183	0.01088	0.00974	....	....	0.01047	
12	0.01878	0.01718	0.01519	....	....	0.01651	
13	0.0295	0.0266	0.0233	....	....	0.0251	
14	0.0464	0.0416	0.0359	....	....	0.0393	
15	0.0734	0.0650	0.0551	....	....	0.0609	
16	0.1162	0.1042	0.0869	0.1172	0.1092	0.0966	0.1089
17	0.1840	0.1613	0.1331	0.1840	0.1705	0.1508	0.1718
18	0.2910	0.2508	0.2008	0.2910	0.2672	0.2326	0.2682
19	0.4560	0.3890	0.3048	0.4565	0.4145	0.3560	0.4165
20	0.7200	0.6008	0.4605	0.7165	0.6430	0.5440	0.6500
21	1.134	0.9240	0.6920	1.123	0.9960	0.8310	1.007
22	1.800	1.515	1.162	1.766	1.545	1.354	1.578
23	2.820	2.320	1.744	2.743	2.370	2.066	2.438
24	4.488	3.557	2.596	4.293	3.642	3.150	3.790
25	7.080	5.440	3.822	6.645	5.570	4.820	5.867
26	11.27	8.300	5.740	10.05	8.510	7.318	9.100
27	17.75	12.52	8.330	15.75	12.89	11.08	13.92
28	28.34	18.90	12.15	24.83	19.54	16.74	21.75
29	44.32	28.05	17.30	37.65	29.08	24.91	33.12
30	70.15	42.08	25.15	58.45	43.75	37.08	50.56
31	110.40	62.45	36.05	89.40	65.08	55.40	77.60
32	172.6	91.45	50.76	134.7	95.40	81.35	116.8
33	279.0	134.0	71.30	208.0	140.5	120.8	179.1
34	433.2	195.6	99.77	309.5	205.8	175.0	265.0
35	684.5	281.8	138.7	459.6	297.3	251.7	396.7
36	1,094.0	405.7	191.6	685.6	429.0	364.2	597.3
37	1,723.0	576.7	263.0	1,014.0	613.5	522.0	887.5
38	2,693.0	817.7	357.0	1,497.0	875.8	735.5	1,294.0
39	4,332.0	1,148.0	480.0	2,193.0	1,235.0	1,048.0	1,927.0
40	6,770.0	1,605.0	650.0	3,202.0	1,736.0	1,461.0	2,791.0

Obviously, then, if the winding volume in cubic inches be multiplied by the resistance per cubic inch, as given in Table VIII, the total resistance of the winding will be known. With this, the total length of wire in the coil may be ascertained from

Table III or from Table VII, and the total weight from Table IX, which gives the weight per 1,000 feet of insulated wire as distinguished from the fourth column of Table III, which gives the corresponding weight of bare wire.

TABLE IX.—WEIGHT IN POUNDS PER 1,000 FEET

Symbol m

B. & S. gage	Enamel	Single cotton	Double cotton	Single silk	Double silk	Cotton enamel	Silk enamel
8	50.55	50.60	51.15	......	......	51.25	
9	40.15	40.15	40.60	......	......	40.70	
10	31.80	31.85	32.18	......	......	32.26	
11	25.25	25.30	25.60	......	......	25.66	
12	20.05	20.10	20.40	......	......	20.48	
13	15.90	15.99	16.20	......	......	16.32	
14	12.60	12.73	12.91	......	......	12.90	
15	10.00	10.10	10.33	......	......	10.27	
16	7.930	8.025	8.210	7.890	7.955	8.180	8.010
17	6.275	6.395	6.540	6.260	6.315	6.480	6.340
18	4 980	5.080	5.235	4.970	5.015	5.160	5.040
19	3.955	4.035	4.220	3.940	3.990	4.120	4.010
20	3.135	3.218	3.373	3.132	3.173	3.275	3.190
21	2.490	2.561	2.685	2.488	2.520	2.625	2.545
22	1.970	2.048	2.168	1.976	2.006	2.118	2.045
23	1.565	1.635	1.727	1.570	1.593	1.668	1.608
24	1.245	1.304	1.398	1.247	1.272	1.335	1.283
25	0.988	1.039	1.129	0.994	1.018	1.071	1.018
26	0.7845	0.8335	0.9140	0.7905	0.8100	0.8570	0.8100
27	0.6220	0.6660	0.7560	0.6280	0.6450	0.6845	0.6445
28	0.4940	0.5325	0.6075	0.4980	0.5140	0.5480	0.5140
29	0.3915	0.4255	0.4890	0.3970	0.4130	0.4375	0.4120
30	0.3105	0.3400	0.3955	0.3160	0.3330	0.3555	0.3295
31	0.2465	0.2762	0.3257	0.2517	0.2678	0.2874	0.2635
32	0.1960	0.2230	0.2700	0.2100	0.2170	0.2348	0.2115
33	0.1550	0.1816	0.2270	0.1611	0.1750	0.1904	0.1683
34	0.1230	0.1478	0.1928	0.1290	0.1412	0.1551	0.1348
35	0.0980	0.1202	0.1600	0.1035	0.1130	0.1286	0.1085
36	0.0776	0.0994	0.1361	0.0823	0.0920	0.1062	0.0872
37	0.0616	0.0822	0.1204	0.0663	0.0740	0.0908	0.0704
38	0.0488	0.0702	0.1049	0.0534	0.0623	0.0778	0.0572
39	0.0387	0.0602	0.0937	0.0424	0.0504	0.0669	0.0463
40	0.0307	0.0519	0.0838	0.0345	0.0429	0.0571	0.0376

A still more direct method of obtaining the weight is to multiply the volume of the coil in cubic inches by the weight per cubic inch, which is given in Table X for each kind of wire.

Calculations similar to the foregoing may be used to determine the size of wire, resistance, number of turns, length and weight

TABLE X.—WEIGHT IN POUNDS PER CUBIC INCH

Symbol w

B. & S. gage	Enamel	Single cotton	Double cotton	Single silk	Double silk	Cotton enamel	Silk enamel
8	0.2540	0.2362	0.2154	......	......	0.2352	
9	0.2411	0.2208	0.1969	......	......	0.2175	
10	0.2382	0.2230	0.2036	......	......	0.2142	
11	0.2381	0.2185	0.1980	......	......	0.2145	
12	0.2374	0.2180	0.1953	......	......	0.2131	
13	0.2350	0.2128	0.1891	......	......	0.2045	
14	0.2314	0.2122	0.1838	......	......	0.2007	
15	0.2308	0.2063	0.1789	......	......	0.1964	
16	0.2301	0.2089	0.1777	0.2318	0.2150	0.1975	0.2178
17	0.2287	0.2042	0.1722	0.2279	0.2130	0.1937	0.2158
18	0.2277	0.2001	0.1648	0.2268	0.2103	0.1860	0.2075
19	0.2262	0.1955	0.1592	0.2240	0.2058	0.1826	0.2062
20	0.2224	0.1912	0.1510	0.2215	0.2015	0.1754	0.2041
21	0.2208	0.1856	0.1454	0.2186	0.1965	0.1706	0.2002
22	0.2198	0.1926	0.1563	0.2167	0.1923	0.1754	0.1973
23	0.2173	0.1869	0.1481	0.2121	0.1858	0.1712	0.1930
24	0.2178	0.1810	0.1414	0.2116	0.1814	0.1641	0.1896
25	0.2165	0.1748	0.1337	0.2040	0.1754	0.1598	0.1849
26	0.2170	0.1697	0.1290	0.1952	0.1693	0.1537	0.1810
27	0.2151	0.1624	0.1225	0.1927	0.1619	0.1479	0.1749
28	0.2160	0.1556	0.1141	0.1915	0.1552	0.1416	0.1725
29	0.2128	0.1461	0.1036	0.1830	0.1468	0.1332	0.1666
30	0.2121	0.1388	0.0966	0.1793	0.1411	0.1281	0.1619
31	0.2097	0.1326	0.0901	0.1729	0.1341	0.1227	0.1574
32	0.2064	0.1244	0.0836	0.1651	0.1257	0.1162	0.1505
33	0.2094	0.1181	0.0784	0.1625	0.1191	0.1113	0.1458
34	0.2045	0.1111	0.0738	0.1532	0.1112	0.1042	0.1372
35	0.2041	0.1032	0.0677	0.1447	0.1023	0.0987	0.1313
36	0.2049	0.0973	0.0630	0.1361	0.0954	0.0933	0.1256
37	0.2032	0.0907	0.0606	0.1286	0.0870	0.0908	0.1196
38	0.1996	0.0870	0.0568	0.1212	0.0812	0.0869	0.1123
39	0.2019	0.0832	0.0541	0.1122	0.0749	0.0847	0.1070
40	0.1985	0.0797	0.0520	0.1017	0.0712	0.0796	0.1003

to meet desired conditions in a given winding space. On the other hand, one may start with a needed number of turns, a needed resistance, and desire to find the size of wire and winding space that will best meet the conditions. In fact, the problem may be to learn any one or more of the unknown variables from any of the others that are known or assumed. The most orderly way of treating these problems is to consider them algebraically.

The factors entering into such problems are of two kinds: those involved in the dimensions of the winding space and those in the characteristics of the wire.

The dimensional factors shown in Fig. 192, or derived therefrom, are as follows:

L = length of winding space.
D = outside diameter of winding.
d = diameter of core over core insulation.
M = mean diameter of winding or diameter of average turn.
T = thickness or depth of winding.
V = volume of winding.

The wire characteristics are

R = total resistance of winding.
r = resistance per cubic inch (Table VIII).
s = resistance per inch of wire (Table VII).
N = total number of turns in coil.
n = number of turns per square inch (Table VI).
W = weight of insulated wire in winding.
w = weight per cubic inch of winding (Table X).
m = weight per 1,000 feet of wire (Table IX).

The following formulæ show some of the most useful relationships among these various factors:

$$T = \frac{D - d}{2} \qquad R = Vr = \pi MLTr = \pi MNs$$

$$M = \frac{D + d}{2} = T + d \qquad N = LTn = \frac{R}{\pi Ms}$$

$$D = \sqrt{\frac{4V + \pi Ld^2}{\pi L}} \qquad W = Vw = \pi MLTw = \frac{\pi MNm}{12{,}000}$$

$$V = \pi MLT = \frac{\pi L(D^2 - d^2)}{4} = \frac{R}{r} = \frac{W}{w}$$

To find the size of wire of a specified kind of insulation that will result in a total required resistance to fill a known winding space, take the size whose value r, in Table VIII, most closely approximates the quotient R/V.

Again, to find the size of wire of specified insulation, to obtain a required number of turns, take the size whose value n, in Table VI, most closely approximates the value N/LT.

An electromagnetic coil will exhibit all the properties of an electromagnet even if wound on a core of wood, or on no core at all. If, however, it has a core of iron, the magnetic properties it displays will be vastly augmented. This is due, as stated in Chap. X, to the greater permeability of iron than that of any of the non-magnetic substances. Hence it is that electromagnets nearly always have iron cores.

The amount of magnetic flux, or, to express it in another way, the total number of lines of force that will be set up in a magnet by a given magnetizing force, will depend on the reluctance of the magnetic circuit formed in part by the core. It must be remembered that the magnetic circuit, even though it consists wholly or in part of air, must always be considered as a *closed circuit.* The lines of force, which we consider as passing through the core and out at one end, always continue as closed loops back into the core at the other end.

Air affords a very much poorer path for magnetic lines than iron. It has very much higher *reluctance.* Consequently, other things being the same, the greater the proportion of iron and the smaller the proportion of air in the length of the magnetic circuit the lower will be the reluctance. In such an electromagnet as is diagrammatically illustrated in Fig. 193, it is obvious that the lines of force, after passing through the iron core, have a long return path through air. The reluctance of the circuit will be comparatively high, even though the iron core be of high permeability.

We may consider the means available for reducing the magnetic reluctance of the core structure with which an electromagnetic coil is associated. The simplest and perhaps the oldest method is merely to bend the core into horseshoe form so that the two poles approach each other, leaving a relatively small air gap for the magnetic lines to traverse. To illustrate this, the straight bar magnet shown at the right of Fig. 194 may be considered as having a continuous winding. For a current in the direction

indicated by the horizontal arrows, the magnetic lines will pass out through the bottom of the core and encounter a long path through air before they return to the other end of the core. If now, this bar, without altering the winding, be bent into horseshoe form, as shown at the left of Fig. 194, the length of the path through air will be greatly shortened. The magnetizing force set up by a given current will not be altered, because the number of ampere turns will remain the same. The reluctance of the magnetic path will, however, be reduced and, as a consequence, the resulting flux will be increased.

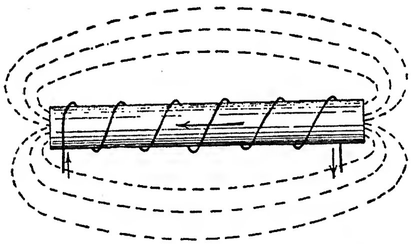


FIG. 193.—Straight-bar electromagnet.

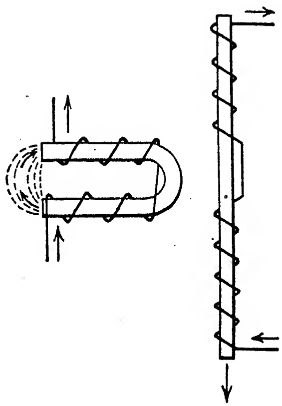


FIG. 194.—Straight-bar and horseshoe electromagnets.

The diagrams of Fig. 194 are useful, also, in showing the relationships between the directions of the windings on the two limbs of a horseshoe magnet. Obviously, this relationship should be the same whether the bar be straight or bent.

A convenient way to remember the relationship between the direction of the magnetizing current and that of the resulting lines of force is illustrated in Fig. 195. If the core be grasped in the right hand, with thumb extended, the lines of force passing through it will be in the direction of the thumb when the current is passing around the bar in the direction of the fingers.

In practice, the horseshoe form of core usually takes the form of a composite structure, such as is indicated in Fig. 196 rather than the simple bent bar of Fig. 194. The presence of the armature, partially bridging the gap between the two poles, further

shortens the air gap with additional lowering of the reluctance. This brings out the fact that the reluctance of the magnetic circuit of any electromagnet depends on the position of the armature. The closer the armature to the poles the lower the reluctance and, consequently, the greater the flux and the stronger the pull.

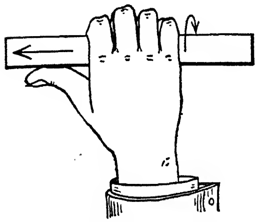


FIG. 195.—Relationship between directions of current and of lines.

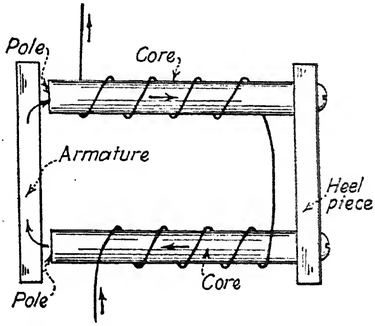


FIG. 196.—Horseshoe electromagnet.

A common example of this form of electromagnet is found in the ordinary two-coil vibrating bell used for door bells in residences, and another in the familiar telegraph sounder.

In many cases, in telephone practice, great economy of space in electromagnetic apparatus is desirable. To further this end, and also that of cheapness, a number of types of magnetic structure employing but a single core and yet affording a path of low reluctance for the lines of force have come into wide use. One of these, shown in Fig. 197, has a straight core which may carry one or more windings, as desired. Each end of this core has an extension projecting laterally a distance just sufficient to clear the winding space. These extensions, or "heel pieces," may be screwed, riveted, or welded to the core ends, or they may be integral with the core and formed by bending over its ends. The magnetic circuit is completed, or nearly so, by the movable armature, hinged at one end to allow for a relatively small movement at the other.

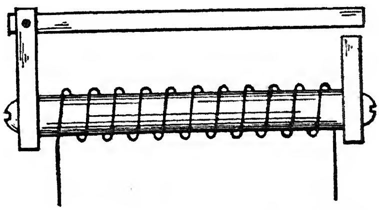


FIG. 197.—Single-core magnet structure.

A similar low-reluctance magnetic structure is shown in Fig. 198. Here, the principal part of the return magnetic circuit is provided by an L-shaped piece, the short end of which is fastened to the rear end of the core. The long end extends parallel to the core and reaches to a point about opposite its free end. The magnetic circuit is closed, or nearly so, by the

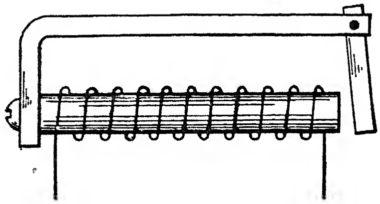


FIG. 198.—Single-core magnet structure.

short iron armature, hinged, in various ways, to the long end of the angle piece to permit a slight movement at the other, which lies opposite the core end. These "single-core" electromagnets of the type shown in Figs. 197 and 198 are very largely used in electromagnetic relays, which form one of the principal elements of central-office switchboard equipment.

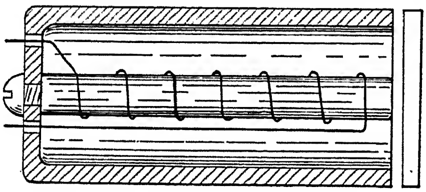


FIG. 199.—Tubular electromagnet.

Another form of low-reluctance magnetic circuit is found in the so-called "tubular" or "ironclad" electromagnet shown in Fig. 199. Here, the core with its helix lies entirely within a tubular shell of soft iron permanently closed at one end. The core is rigidly attached at one end to the closed end of the shell. The armature, a disc of soft iron, closes—or nearly closes—the magnetic circuit between core and shell at the other end. The lines of force pass from one end of the core into the shell, returning, through it, the armature and the short air gaps, to the other end of the core.

This tubular form of electromagnet has the advantage of affording a magnetic path of very low reluctance. Consequently, its attractive force, or holding force, is very great in proportion to the amount of magnetizing force applied.

Another advantage of the tubular form of magnet is that it possesses practically no "stray field." In an ordinary bar magnet the returning lines of force spread out into the air, extending considerable distances from the core. Theoretically, there is no limit to this spreading out of the field. Practically, however, with an ordinary bar magnet, the magnetic field in the space surrounding the core is often strong enough to be observed, and sometimes to produce annoying results, at distances of several feet. The annoying results from this cause are manifested by the presence of induced currents in neighboring coils which are associated with circuits carrying voice currents. Such induction, from one circuit to another, results in "cross-talk" where voice currents are involved. In the tubular magnet the return path through the iron shell is so complete that nearly all of the returning lines of force pass through it, stray field being practically eliminated. Such coils may, therefore, be grouped together so closely as to be in actual contact, side by side, without causing trouble due to induction from one to the other.

All of the foregoing types of core structures are available for such electromagnetic devices as relays, drops and bells, in which definite mechanical movements are required of armatures to carry out the desired function. These general types are also used as the cores of impedance coils, transformers, induction coils and other electromagnetic coil devices which need no movable armatures because the production of mechanical movement forms no part of their functioning.

Another type of core structure, definitely limited to use with the latter type of device requiring no mechanical movements, is that having permanently closed magnetic circuits of fixed characteristics. Perhaps the most perfect achievement in this direction is to be found in the cores of the so-called "toroidal types" of coil, developed for use where low and fixed conditions of reluctance are desired. These cores consist of closed circular rings and, hence, present no air gaps.

Impedance coils and repeating coils wound on such cores are typically represented as in Fig. 200. Here, the representation

is of a repeating coil or transformer. In practice the windings extend entirely around the coil and nearly fill the central opening. Two such coils as actually manufactured are shown in Fig. 201.

Obviously, whatever magnetic field is set up by one of the windings will pass through the other winding. Any variations in one will cause corresponding currents to flow in the other. Owing to the symmetrical distribution of the windings and the completely closed low-reluctance magnetic circuit, there is little tendency toward stray field, thus largely avoiding crosstalk troubles in neighboring circuits.

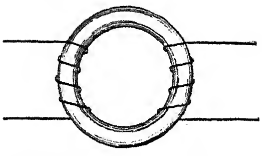


Fig. 200.—Toroidal type of coil.

The toroidal type of coil is widely used for repeating coils, transformers and impedance coils. Its principal application, however, is for loading coils for improving the transmission qualities of telephone lines, for which use it was first extensively developed.

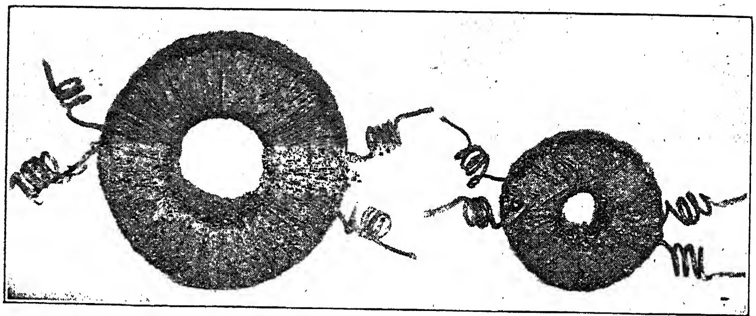


Fig. 201.—Toroidal transformer. (*Courtesy of Bell Telephone Laboratories.*)

One cause of energy losses in electromagnetic devices is eddy currents. A varying current in the winding will tend to produce other currents by induction in all near-by conductors. One of the near-by conductors is, of course, the core itself. When, therefore, the current of any electromagnetic coil is varying, it causes induced currents to circulate within the conducting material of the core. For the most part, these eddy currents are useless and often actually harmful.

In most direct-current electromagnets, the elimination of eddy currents is not important. If, however, the magnet is required

to attract or release its armature with great promptness, then eddy currents become objectionable, since they always tend toward delaying action. In other electromagnetic devices of which induction coils, transformers and impedance coils are examples, it is desirable to suppress them, since they consume a part of the energy that should be employed in useful directions.

There are two common ways of minimizing eddy currents in cores, each of which depends on the principle of increasing the ohmic resistance of the path through the core that the eddy currents must follow. One of these ways is to use a grade of magnetic material that has inherently a high electrical resistivity and, at the same time, satisfactory magnetic properties. As pointed out in Chap. X, silicon steel is such a material, because its electrical resistivity is about five times that of ordinary magnetic iron, while its magnetic permeability is higher throughout the range of flux densities most used in speech-transmission apparatus. The other way is to subdivide the core material so as to increase the electrical resistance of the path followed by the eddy currents. Hence it is that in induction coils, transformers, and other devices that are used with rapidly varying currents, such as voice currents, and required to work at high efficiency, the cores are usually built up of small wires or thin sheets instead of being solid. Hence it is, also, that such magnetic materials as silicon steel, having high permeability and high resistivity, are largely used in such cores instead of the ordinary magnetic iron.

As examples, the ordinary straight-core telephone induction coil is usually wound on a core formed of a bundle of soft iron wires, or, in a later practice of the Bell companies, on a bundle of thin, straight, flat strips of sheet steel. During recent years the practice in the construction of certain types of telephone transformers has tended toward the building up of cores from flat steel punchings, a method long employed in power and lighting transformers. A modern telephone transformer with a core of this kind is shown in Fig. 202. This design is instructive as showing how form-wound coils of the type already referred to in Fig. 190 are being used in telephone work.

Until recent years the cores of all toroidal coils, whether used for loading or other purposes, were formed by winding iron wire on a round mandrel until enough turns had been applied to give a ring of the required cross-section. This form of con-

struction is still used for some purposes, but for loading coils it has been supplanted in all new work by the dust core described in Chap. X. The cores so formed offer low reluctance and at the same time high electrical resistance. The latter attribute reduces eddy-current losses to an almost imperceptible value. Also—and of great importance—a magnetic stability hitherto not secured in other forms of core material is attained. It seems probable that the use of dust cores will find wide application in other fields of telephone work.

The use, as core materials, of permalloy and other iron-nickel-cobalt alloys, described in Chap. X, is still in its infancy

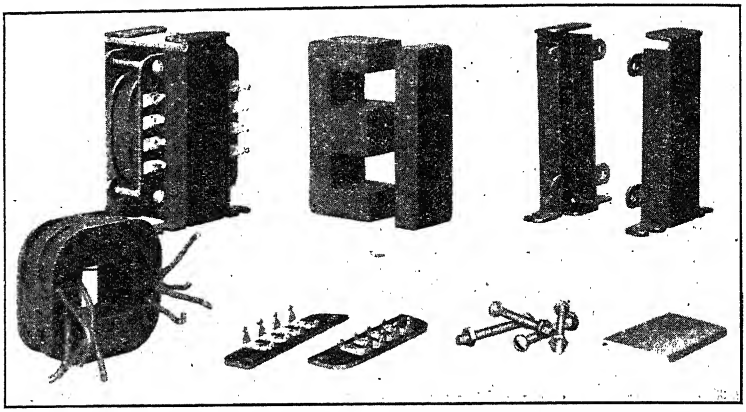


FIG. 202.—Modern input transformer. (*Courtesy of Bell Telephone Laboratories.*)

but is of great importance. Permalloy is being employed in the magnetic structures of certain telephone receivers and also in relays where increased sensitiveness is required. It is also being used in the cores of some amplifier transformers for the purpose of improving the efficiency at the lower range of frequencies and thus helping to make the amplification uniform throughout the entire range of voice frequencies. Transformers mechanically like that of Fig. 202 are employing cores of permalloy with this object in view.

If a core has two windings that are exactly alike in all respects and are exactly alike in their mutual relations to the core, then it is readily seen that if equal currents be sent through them in opposite directions, no magnetism will be produced. The magnetic effects of the two windings, being equal and

opposite, exactly neutralize each other. Such windings are termed "differential" and they play a rather important part in telephone systems.

The principle of differential windings may be illustrated in connection with Fig. 203. Two exactly similar wires are wound on the core side by side and very close together throughout their lengths. We may assume that they are so close together as to occupy the same physical space and that they are joined together at *b*, the other ends, *a* and *c*, remaining separate. If current passes from *a* to *b* through one winding, magnetic flux in the direction of the horizontal arrow will be set up in the core. The same will be true and flux in the same direction will occur if current passed through the other winding from *c* to *b*. If, however, current passed from *a* to *c*, thus traversing the two wind-

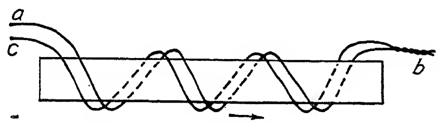


FIG. 203.—Differential winding.

ings in opposite directions, the flux which one winding tends to set up will be exactly neutralized by that of the other, and no resulting magnetism will occur.

While complete neutralization is easily conceivable, it is not so easy to realize in practice, because the two wires cannot simultaneously occupy the same physical space and, therefore, cannot have exactly the same relationship to the core. It may be closely approximated, however, by first twisting the two wires together and then winding them on the core. This procedure, while resulting in a magnet that is sufficiently differential for practical purposes, has some structural objections. It is difficult to maintain proper insulation between the two windings, where insulation and space requirements are exacting. Moreover, owing to the close proximity of the two wires throughout their length, the electrostatic capacity between them is increased, and this in some cases is objectionable. Because of these features, methods of winding that secure less exact differential results are often employed.

One, indicated in Fig. 204, is to place one winding inside the other, each extending throughout the length of the core. This affords a rather poor approximation to the desired result.

If the two windings have the same number of turns, necessary for neutrality, then the outer has higher resistance than the inner. If the resistances are made equal, the outer winding has fewer turns than the inner. Also, the different relationship of the two windings with respect to the core makes it impossible for them to exert exactly equal and opposite effects on it.

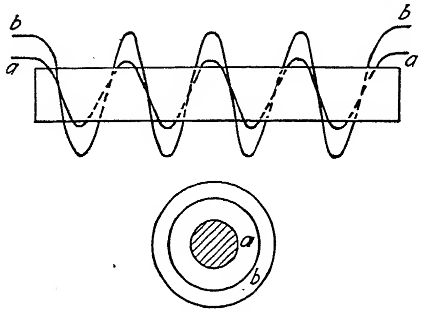


FIG. 204.—Concentric windings.

A much closer approximation to differentiality may be attained by "sandwiching" the concentric windings, that is, by first putting on half the turns of winding a, then all the turns of winding b, and over these the remaining half of winding a, as shown in Fig. 205. In this way the desired equality both as to number of turns and resistance may be secured. Also, the

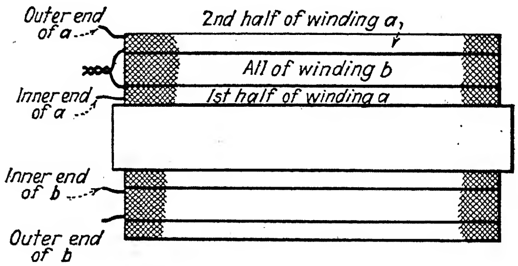


FIG. 205.—"Sandwich" concentric windings.

desired equal and opposite magnetizing effects of the two windings are approached but not fully realized. In such a coil, however, a high degree of insulation between the two windings may be maintained, and the capacity effects between them are not so pronounced.

Another plan of winding is to place equal and opposite coils on the respective half-lengths of the core, as shown in Fig. 206.

This is called "tandem winding." It has its uses, but it is only partially differential with respect to direct currents, and scarcely differential at all with respect to voice currents. The reason one winding will not directly neutralize the other, even with direct currents, is that all of the magnetic lines of force set up by one of the coils do not thread through the other coil. Some of them leak away and return through the air without passing to the distant end of the core. For voice currents this effect is so pronounced that the two tandem coils act almost as if they were on separate cores, and there is little differential effect.

The differential winding formed by twisting the two wires together before wrapping them on the core (Fig. 203) will so completely neutralize each other with respect to voice currents

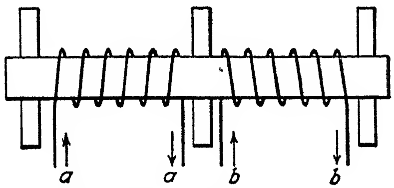


Fig. 206.—Tandem windings.

as to present practically no impedance other than that of the pure resistance of the wire. This is one of the common ways of producing non-inductive resistances. Such a condition is only fairly approximated in the sandwiched concentric differential windings of Fig. 205.

The making of electromagnets that act with varying degrees of rapidity has assumed great importance in telephone switching. This is particularly true of the electromagnets employed in some of the relays used in automatic or machine-switching systems. The time difference between the responses of "slow-acting" and "quick-acting" relays has become one of the absolutely essential elements of automatic switchboard practice. The use of relays that would discriminate with certainty in the time of their response marked the passing of the old three-wire automatic system and made possible the two-wire system which superseded it.

The ordinary straight bar magnet, such as is diagrammatically shown in Fig. 193, may be made to attract and release its armature quite rapidly in response to a succession of current impulses.

The response may be made even more rapid if the core is formed of a bundle of soft iron wires rather than a solid bar. On the other hand, if the core be surrounded by a copper sleeve under the winding or by a heavy copper slug at either end of the winding, the magnet will be found to have become sluggish in its action, particularly with respect to release after the cessa-

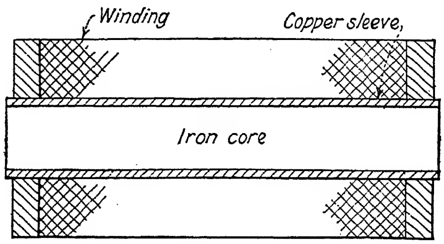


FIG. 207.—Copper sleeve for slow action.

tion of the current. The reason for these differences in speed is found in the varying degree to which induced currents are allowed to flow in the nearby conductors, such as the core or the copper sleeves.

The induced currents, whether in core, copper sleeve, or other adjacent conductors, are always of such a direction as to tend to prevent change in the existing magnetic flux. Thus

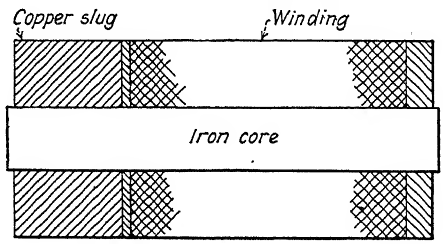


FIG. 208.—Copper slug for slow action.

it is that the substitution of a bundle of iron wires for a solid core will tend to increase the speed of the magnet, since it reduces the eddy currents in the core. On the contrary, the copper sleeve or slug causes the magnet to act slowly because of the currents that are induced in them. It is for this reason that the tubular magnet of Fig. 199 is sluggish in action. The iron shell affords an excellent opportunity for the flow of eddy currents.

Magnets purposely rendered slow acting by the application to their cores of copper sleeves and slugs are shown in Figs. 207, 208 and 209. The heavy copper slugs may vary in length according to the degree of sluggishness desired. They are ordinarily made to extend from one-quarter to one-half the length of the core. Some variation as between the speed of pulling up and that of release may be made by putting the slug at the heel end or at the armature end of the core. If sluggish release

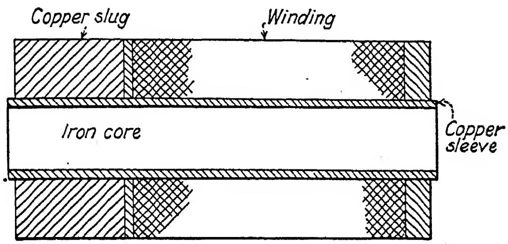


FIG. 209.—Copper sleeve and slug for slow action.

is desired, with little effect on the speed of attraction, the slug is put at the heel end. If it is desired to retard the pull up as well as the release, then the slug is put at the armature end. A somewhat greater range of timing may be had by variously combining a sleeve and a slug on the same core, as shown in Fig. 209.

In general, however, electromagnets employed in relays and especially designed to be sluggish are only slightly slower in the pull up. The principal effect is the slowing down of the release of the armature, this being delayed in ordinary relay practice by a considerable fraction of a second.

CHAPTER XIII

RESISTORS

Coils of the type generally considered in the last chapter offer more obstruction to rapidly varying currents than to unvarying currents. For steady currents they present only the ohmic resistance of the wire, while for varying currents there is the added impedance due to the inductance of the coil. The difference is often very great. An exception to this is found in coils such as that of Fig. 203, in which, by carefully balanced double windings, the inductive effects of one winding are neutralized by those of the other. Such non-inductive coils present only the ordinary resistance of the wire, their impedance being practically alike for both steady and varying currents.

The fact that inductive coils offer greater impedance to varying currents than to steady currents, while non-inductive coils behave alike to each, is turned to many useful purposes in telephony. While the inductive coil is outstandingly the more important of the two, the non-inductive coil, or, broader yet, the non-inductive resistance, deserves attention in this chapter.

We may refer to that class of resistance devices which behave substantially alike to both alternating and direct currents as resistors. They rely on ordinary resistance rather than on inductive effects to limit the flow of current. They are ordinarily referred to in telephony as non-inductive resistances, a term that is apt but not always accurate. Generally speaking, many of the devices, so referred to, do, in fact, have some slight electromagnetic inertia, though this may be practically negligible for the purposes under consideration. A case in point is the ordinary incandescent lamp in which the filament is looped one or more times. These lamps, as will be shown, are extensively employed in telephony as resistors and are commonly classed as non-inductive resistances, although they do offer some slight additional impedance to varying currents.

The uses of the resistor or non-inductive resistance are those which grow out of the relationship between current, electro-

motive force and resistance, as expressed by the familiar equations of Ohm's law, $I = E/R$, $E = IR$ and $R = E/I$. These uses in telephony are to limit current flows, to establish differences of potential at definite points in a circuit, to act as by-passes or shunts for voice currents around paths containing objectionable impedances, to afford definite leakage paths from one circuit to another, to effect adjustments of current or electromotive force for various purposes, to develop heat at desired points and to act as standards of resistance with which to compare other resistances.

As many of the resistors used in telephony are formed of wire, the characteristics of wires available for this purpose deserve consideration. The comparative resistances of various metal conductors with respect to pure copper have already been given in Table I.

Copper wire could be and often is used. Of course, any desired resistance may be obtained by employing a sufficient length of a given size of it, but, where resistance alone is the desideratum, copper is a poor material. It is too good a conductor, which means that the resulting structure, for a given resistance, would often be unduly bulky and costly, particularly if considerable current-carrying capacity was required. Moreover, if the copper wire is wound in the ordinary way, its greater length for a given resistance results in a comparatively large number of turns, with consequently larger inductance. This may be avoided by adopting some sort of non-inductive winding, which, in itself, is costly and may be otherwise objectionable. Again, copper has a comparatively low melting point, objectionable where considerable heat is likely to be developed; and, lastly, it has a comparatively high temperature coefficient with respect to resistance which may or may not be objectionable, according to the uses to which the resistor is to be put.

The temperature coefficient of a conductor relates to its change in resistance with change in its temperature. It is usually expressed as a factor by which the resistance of a conductor at a given temperature is to be multiplied to determine the change in resistance due to a rise in temperature of one degree.

All pure metals have the property of increasing their resistance as their temperatures rise. These are said to have positive temperature coefficients. On the other hand, some substances, of which carbon is a notable example, decrease in resistance

with increase of temperature. These have negative temperature coefficients.

In order to compare the resistivity of different materials, it is necessary to compare the resistances offered by them in paths of exactly equal lengths and cross-sections. In this way dimensional differences that would affect resistance are eliminated and the comparison is resolved down to the *specific resistances* of the materials themselves. For this comparison the resistivity is often expressed in terms of the resistance through one cubic centimeter of the material from one face to the opposite face.

Such resistances of various metals at a temperature of 0° C., together with the mean temperature coefficients of each, are shown in Table XI.[1]

It has been found that the presence of impurities in metals cause pronounced variations in both the resistivity and the temperature coefficient. This in part accounts for the variations that will be noticed in different tables giving these properties

TABLE XI.—ELECTRIC RESISTIVITY AND TEMPERATURE COEFFICIENTS

Metal	Resistance at 0° C. per centimeter cube in C.G.S. units*	Mean temperature coefficient between 0° C. and 100° C.
Silver, electrolytic and well annealed	1,468	0.00400
Copper, electrolytic and well annealed	1,561	0.00428
Gold, annealed	2,197	0.00377
Aluminum, annealed	2,665	0.00435
Magnesium, pressed	4,355	0.00381
Zinc	5,751	0.00406
Nickel, electrolytic	6,935	0.00618
Iron, annealed	9,065	0.00625
Cadmium	10,023	0.00419
Palladium	10,219	0.00354
Platinum, annealed	10,917	0.00366
Tin, pressed	13,048	0.00440
Thallium, pressed	17,633	0.00398
Lead, pressed	20,380	0.00411
Bismuth, electrolytic	110,000	0.00433

* The C.G.S. unit of resistance is one thousand millionth part of an ohm.

[1] As determined by Fleming and Dewar, *Philosophical Magazine*, September, 1893; see also "Encyclopedia Britannica," 11th ed., vol. 6, p. 856.

of metals. These variations are particularly noticeable in alloys. Another cause of variation in pure metals as well as alloys is the degree of hardness. This, in wires of identical composition, will depend principally on the degree to which the wire has been annealed between the various stages, or after the completion, of the wire drawing.

By variously compounding different substances wide ranges of conductivity, and its reciprocal resistivity, may be obtained; and also wide variations in temperature coefficients extending from positive through zero to negative.

Obviously, this property by which the resistance of a conductor varies with its temperature is one of the things that must be considered in choosing a resistance material. Sometimes, it may be made to serve useful purposes, in many cases it may be practically ignored and in others it must be avoided as far as possible.

For certain purposes wires having a zero temperature coefficient have been greatly sought. As an example, in standard resistance coils it is obviously desirable to have a wire the resistance of which will not change throughout the range of temperature to which it is to be subjected. Another desired characteristic is that the wire shall not oxidize or otherwise change its chemical or physical structure through long periods of time, and within the ranges of temperature likely to be encountered. Still another is that it shall have a melting point high enough not to be fused under current-carrying conditions likely to be imposed.

Probably the most common resistance wire for telephone purposes is an alloy called "German silver" or "nickel silver." It contains no silver, being composed of copper, nickel and zinc. This is ordinarily made in two grades known, respectively, as 18 per cent and 30 per cent; these percentages referring to the amount of nickel in the alloy.

The two grades, 18 and 30 per cent, have, respectively, resistances about 18.2 and 28 times the resistance of corresponding gages of annealed copper wire. The temperature coefficients of the two grades are, respectively, about 0.00016 and 0.00024 per degree Fahrenheit.

The various characteristics of weight, length and resistance of 18 per cent German silver wire are shown in Table XII. For the 30 per cent grade, the weights and lengths, size for size,

Table XII.—18 Per Cent Nickel-silver Wire

B. & S. gage	Weight, lb. per 1,000 ft.	Length, ft. per lb.	Resistance at 75° F.	
			Ohms per foot	Ohms per lb.
16	7.613	131.3	0.07318	9.612
17	6.040	165.6	0.09227	15.27
18	4.790	208.8	0.1164	24.29
19	3.799	263.2	0.1467	38.62
20	3.011	332.1	0.1850	61.46
21	2.389	418.6	0.2333	97.69
22	1.894	527.9	0.2941	155.3
23	1.502	665.1	0.3710	247.0
24	1.191	839.6	0.4678	392.7
25	0.9449	1,058.0	0.5899	624.3
26	0.7493	1,333.0	0.7438	992.7
27	0.5943	1,683.0	0.9386	1,579.0
28	0.4711	2,123.0	1.183	2,511.0
29	0.3735	2,677.0	1.491	3,991.0
30	0.2962	3,376.0	1.879	6,343.0
31	0.2350	4,255.0	2.371	10,090.0
32	0.1864	5,364.0	2.990	16,046.0
33	0.1478	6,766.0	3.771	25,510.0
34	0.1172	8,532.0	4.756	40,577.0
35	0.09295	10,758.0	5.997	64,515.0
36	0.07369	13,570.0	7.560	102,590.0
37	0.05845	17,109.0	9.532	163,040.0
38	0.04636	21,570.0	12.020	259,300.0
39	0.03675	27,211.0	15.160	412,450.0
40	0.02917	34,282.0	19.110	655,070.0

Specific gravity 8.68.
Resistance 189 ohms per circular mil foot at 75° F.
Temperature coefficient 0.00016 per degree Fahrenheit.
Resistance approximately 18.2 times that of copper.
The resistance is subject to a variation of from 5 per cent below to 10 per cent above the figure given.

are practically the same as in this table. The resistances, however, whether in ohms per foot or per pound, run about 53 per cent higher for the 30 per cent than for the 18 per cent grade.

Another resistance wire, known by the trade name of 1A1A, is a copper nickel alloy. It has an extremely low temperature coefficient (about 0.000003 per degree Fahrenheit) which for

most purposes is negligible. In this respect it satisfies requirements for use in measuring instruments or for other purposes where constancy of resistance under varying temperature is desired. It has a resistance about 28 times that of copper and does not easily corrode. Its melting point is about 2250° F.

Nickel-steel resistance wire has a resistance about 50 times that of copper and about 2.7 times that of 18 per cent German silver wire. It is one of the cheapest wires for use where resistance is the only consideration. It should not be used for temperatures exceeding 800° F. Also it has a tendency to rust and scale which makes its use generally undesirable in sizes smaller than about No. 20 B. & S. gage.

Nichrome wire, made of a patented alloy of nickel, chromium and iron, is remarkable on account of being practically non-corrosive in air for all temperatures up to its melting point. This is high, about 2820° F. Because of these characteristics it is widely used for electric-heating devices for domestic and industrial purposes. It forms the heating element of most soldering irons, which are largely used in telephone work. Its temperature coefficient is in the neighborhood of 0.00011 per degree Fahrenheit.

It is often necessary to consider the current-carrying capacity of a resistance wire in order to adapt it safely to the purposes required of it. It is to be remembered that the current-carrying capacity of a conductor depends very much on its environment, as affecting the rapidity with which the heat generated in it may be dissipated by conduction, radiation or convection. Any tables attempting to show the carrying capacity of wire must, therefore, be used with discretion, owing to the wide variations that may be caused by the immediate surroundings of the conductor, particularly with respect to the freedom with which air or other cooling medium may circulate about it. Such tables must be taken only for general guidance, final determination always depending on results of actual experiment.

Table XIII gives the current-carrying capacity for 18 per cent German silver wire coiled in air with free radiation. The currents, given in amperes, are such as will cause a rise in temperature of from 100 to 600° F. as noted at the top of each column, above an initial temperature of 75° F.

Similar current-carrying capacities for corresponding rises in temperature may be determined for other kinds of wire, used

TABLE XIII.—CURRENT CARRYING CAPACITY OF 18 PER CENT NICKEL-SILVER WIRE

B. & S. gage	Bare wire coiled in air with free radiation — Rise in temperature above 75° F. (24° C.)					
	100° F.	200° F.	300° F.	400° F.	500° F.	600° F.
16	6.6	9.3	11.5	13.2	14.8	16.2
17	5.5	7.8	9.5	11.0	12.3	13.5
18	4.7	6.6	8.1	9.4	10.5	11.5
19	3.8	5.4	6.6	7.6	8.5	9.3
20	3.3	4.7	5.7	6.6	7.4	8.1
21	2.8	4.0	4.9	5.6	6.3	6.9
22	2.3	3.3	4.0	4.6	5.1	5.6
23	1.95	2.8	3.4	3.9	4.4	4.8
24	1.63	2.3	2.8	3.3	3.7	4.0
25	1.39	2.0	2.4	2.8	3.1	3.4
26	1.17	1.7	2.0	2.3	2.6	2.9
27	1.00	1.4	1.7	2.0	2.2	2.5
28	0.83	1.0	1.5	1.7	1.9	2.0
29	0.73	0.9	1.3	1.5	1.6	1.8
30	0.58	0.82	1.0	1.2	1.3	1.4
31	0.49	0.69	0.85	0.98	1.1	1.2
32	0.41	0.58	0.71	0.82	0.92	1.0
33	0.35	0.50	0.61	0.70	0.78	0.86
34	0.29	0.41	0.50	0.58	0.65	0.71
35	0.25	0.35	0.43	0.49	0.55	0.60
36	0.21	0.29	0.36	0.41	0.46	0.50
37	0.17	0.25	0.30	0.35	0.39	0.43
38	0.14	0.20	0.25	0.29	0.32	0.35
39	0.125	0.18	0.22	0.25	0.28	0.31
40	0.105	0.15	0.19	0.21	0.24	0.26

under the same conditions by multiplying the values of Table XIII by the following factors:

Copper, 4.25.

30 per cent German silver, 0.8.

Nickel-steel, 0.607.

1A1A wire, 0.795.

With due regard to the amount of heat to be developed, resistance wire may be insulated with any of the coverings employed in magnet wire or with asbestos or it may be used

bare. For the coils of measuring instruments and many of those used in telephony, where only small currents are to be carried, silk, cotton or enamel coverings are used, the same considerations governing as in magnet wire. If, however, they are to carry large enough currents to develop much heat, silk and cotton must be avoided; and enamel, also, if the temperatures are to run above about 400° F.

For higher temperatures than this asbestos covering may be used, but it is bulky and some sort of bare-wire winding is usually found most effective. With bare wire the insulation between successive layers and also between the core and the first layer is usually composed of mica sheets. The insulation between adjacent turns in the same layer is usually secured either by so spacing the wires that they will not come into actual contact or by applying an intervening thread of asbestos.

Most resistors, or resistance units, or non-inductive resistances, as they are variously called, are made up by winding the proper length of the required size and kind of resistance wire on a spool or card, so that they may be conveniently handled or mounted in association with other apparatus. A common way is to use an ordinary spool, the finished resistor appearing in some such form as shown in Fig. 191. Often a brass core is used to aid in heat dissipation. For the closest approximation to non-inductiveness the wire is doubled and twisted, and the two inner ends soldered to the two terminals. After winding the two outer ends are soldered together, the resulting winding being of the type diagrammatically shown in Fig. 203. Spool-wound resistors of this type, with little or no special provision for the dissipation of heat, must be used with care in this respect. Sometimes they are employed where the creation of heat is the object in view.

A better form of resistor for many telephone purposes has been developed by the Western Electric Company for the Bell operating companies and is now generally used also by other concerns. This is shown in Fig. 210. It is wound of bare resistance wire on a card of mica or micanite, the turns being so spaced along the card as to avoid contact with each other. The winding is then protected by two outer cards of mica which are held in place by metal clips tightly clamped to their edges. Screw-threaded posts, projecting from these clips, serve to secure these units to their mounting plates. These posts also form the

terminals of the units and are provided with soldering tips for attaching the wires. The overall dimensions of these resistors are length: $4\frac{11}{16}$ inches, width, $1\frac{1}{16}$ inch, and thickness, $\frac{3}{8}$ inch. They have been standardized in resistances from a fraction of an ohm up to 9,000 ohms or more, according to the requirements of the circuit in which they are to be used. Ordinarily, their resistances do not vary more than 5 per cent, plus or minus, from the specified rating, but variations may be confined to closer limits if desired. In some cases the center point of the resistance is brought out to a third terminal between the two end terminals, thus conveniently affording pairs of matched resistances in a single unit where the circuit requirements make this desirable.

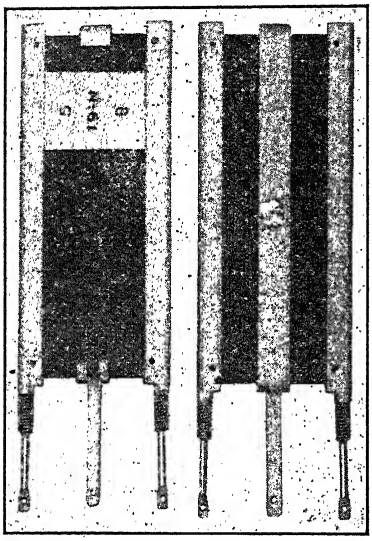


Fig. 210.—Mica-card resistance unit. (*Courtesy of Bell Telephone Laboratories.*)

These mica-card resistances are usually mounted vertically with some air space between them so as to afford free circulation for air-convection currents. In switchboard work they are usually carried on mounting plates like those upon which relays and other elements of apparatus are carried. This permits them being brought into convenient wiring relationship with their associated apparatus. They have been designed with particular regard to freedom from fire hazard.

A type of resistor developed by H. Ward Leonard, and now manufactured by the company bearing his name, employs a resistance wire embedded in vitreous enamel. This is widely use in general power and lighting work and is finding increasing use in telephony, in cases where the question of heat dissipation cannot be satisfactorily met by the ordinary resistors so far considered. These resistors are made by winding a resistance wire of low electrical temperature coefficient on a porcelain-like tube. It is then coated with a powdered vitreous enamel and fired to a red heat until the enamel fuses to the wire and the tube.

The unit thus becomes a solidified mass, in which the resistance material is so embedded as to be protected mechanically, electrically and chemically. The close union between the vitreous enamel and the surface of the wire permits the heat to be rapidly conducted away from the wire and carried off by convection and radiation from the outer surface of the whole mass. One of these resistors is shown wound ready for the enameling process

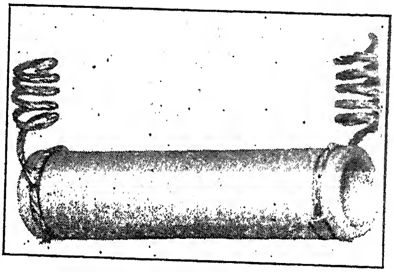


Fig. 211.—Vitreous enamel resistor before enameling. (*Courtesy of Ward Leonard Electric Co.*)

in Fig. 211 and after completion in Fig. 212. Here the terminals are of the pigtail type of braided copper conductors. The joints between the terminals and resistance wire are included within the enameled mass.

Various other designs of small vitreous enameled resistors have been standardized, by the Ward Leonard Company, for

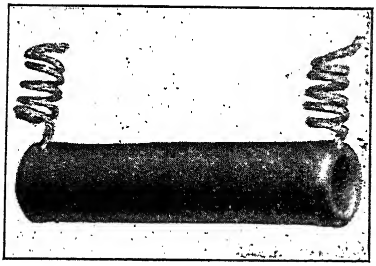


Fig. 212.—Vitreous enamel resistor. (*Courtesy of Ward Leonard Electric Co.*)

telephone use. One of these has soldering terminals at the ends of screw-threaded posts, the posts forming the means of attachment to the mounting plates. This permits these resistors to be mounted interchangably with those of the mica-card type of Fig. 210. Another type (Fig. 213) is adapted for Edison lamp-socket mounting.

In connection with vacuum tube circuits small units of very high resistance—usually expressed in megohms—are required. The external appearance of a common form of these is shown in Fig. 214. The resistance is enclosed within a glass tube with a brass cap at each end. The length of this unit is about 2 inches and the external diameter of the terminal caps 5⁄16 inch. The resistance element within has been formed in a number of ways, as by a mark of india ink on a strip of paper, by a strip of card-

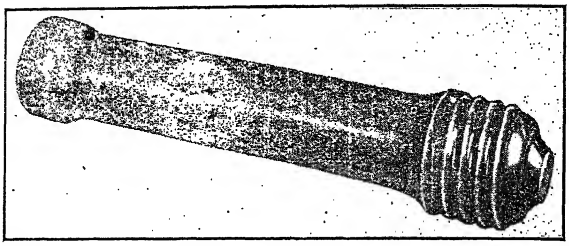


Fig. 213.—Vitreous enamel resistor. (*Courtesy of Ward Leonard Electric Co.*)

board saturated in a carbon solution, by a column of carbon or carbon compound within the tube and by a very thin coating of metallic conductor on the inner surface of the tube.

Ordinary commercial incandescent lamps make effective resistor units for telephone work in cases where their bulk is not objectionable. An instance of such use is found in the common leads for conveying ringing or battery current to groups of switchboard apparatus. For such purposes, commercial lamps

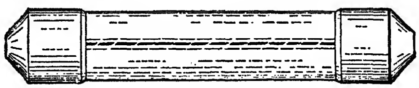


Fig. 214.—Tubular high-resistance unit.

have a number of advantages; they carry comparatively large currents, they directly indicate by their glow whether current is flowing, and whether it is abnormal, and they are safe from the standpoint of fire hazard.

For temporary work, as in emergencies, banks of commercial incandescent lamps afford effective readily adjustable resistors for handling currents of moderate volume and voltage. They are cheap, quickly obtainable and easily mounted and wired. By variously combining them in series and multiple relationship, a wide range of resistance and current capacity may be obtained.

Sometimes resistors with pronounced temperature coefficient characteristics in one direction or the other are desired. For such purposes incandescent lamps are often adaptable. When a resistor is wanted that becomes a better conductor for larger currents, the old-fashioned carbon incandescent lamp may be used on account of the negative temperature coefficient of its filament. On the other hand, the tungsten lamp with its positive coefficient acts in the reverse direction.

On account of the high positive temperature coefficient of iron wire, it has been proposed to employ coils of it, enclosed in exhausted glass tubes, as resistors. These, called "ballast coils," tend to minimize changes in the current flowing through a circuit by increasing resistance for larger currents and lowering it for smaller. This effect is enhanced by employing low-pressure hydrogen in the bulb as in the modern ballast lamp.

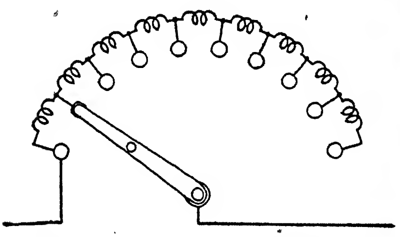


Fig. 215.—Radial-arm type rheostat.

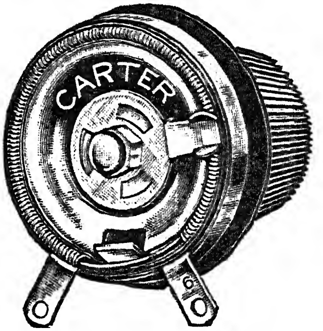


Fig. 216.—Small spiral-wire rheostat. (*Courtesy of Carter Radio Co.*)

The term "rheostat" is applied to a readily adjustable resistance. Usually it consists of a number of resistors, or resistance units, in connection with some form of switching device by which they may be successively cut into or out of circuit. A common form is the radial-arm type shown diagrammatically in Fig. 215. The individual resistors of such a rheostat may be of any of the forms heretofore considered, provided they have the required electrical and heat-resisting characteristics.

Another way of making the resistor in radial-arm or other forms of sliding-contact rheostats, is to wind it as a continuous length into a long spiral and so disposing this spiral that the moving contact will sweep over its successive convolutions. Such a rheostat is shown in Fig. 216. Here the long spiral coil

of the resistor is bent into circular form, so as to be engaged by the radial arm which, in this case, is turned by a knob. Rheostats of this type are used for such purposes as controlling the filament current of vacuum tubes. They have the advantage of affording a large number of small resistance steps in a simple and inexpensive manner.

Another form of rheostat, common in the older types of resistance-measuring instruments, consists of one or more rows of insulated brass blocks, between the successive pairs of which resistance units of definite known values are connected. By inserting metallic plugs between the blocks, the corresponding units are short circuited, and their resistance thus removed from the circuit. This is known as the "plug type" rheostat.

All of the foregoing types of rheostats vary the resistance by steps. The resistor units comprising these steps may be any of the types heretofore referred to. If of wire, they are non-inductively or inductively wound, according to whether the added impedance due to inductance is objectionable or not for the purposes under consideration. The straight-carbon resistors are in themselves non-inductive and the incandescent lamp resistors nearly so.

Of the resistors and rheostats employing other resistance material than metallic wire, those of carbon are most important. Carbon has high specific resistivity, withstands high temperatures and does not readily oxidize. The fact that it has a negative temperature coefficient may or may not be advantageous, depending on the use. The fact that the resistance of the contact between its adjacent surfaces varies so markedly with varying pressure makes it particularly useful where varying resistances are required.

Probably the most common resistor used in telephone practice is the carbon button of the transmitter discussed elsewhere. Of recent years, however, this simple principle of varying contact resistance with varying pressure has been made available for a wider range of usefulness. Its adaptability extends from the control of filament current in small vacuum tubes up to heavy duty motor control in hoisting operations and rolling mill work. The outstanding practical development in this line is due to Mr. Lynde Bradley of the Allen-Bradley Company of Milwaukee.

The principal resistor element of this company consists of a pile of thin graphite discs enclosed in a porcelain-lined steel tube

in such manner as to largely exclude air. The terminals of the pile consist of metal plates making contact with the end disks respectively. These have shanks extending through the ends of the tube in such manner as to insulate them from it. Either by a screw and hand-wheel arrangement or by some sort of lever action, the pile of carbon disks may be subjected to a wide variation in pressure.

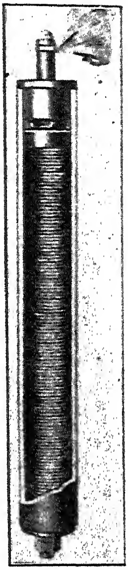


Fig. 217.—Carbon-disk rheostat unit. (*Courtesy of Allen-Bradley Co.*)

When the column of discs is under no compression the contact resistance between successive discs is comparatively high. As the pressure on the column is increased, the resistance gradually decreases until finally most of the contact resistance disappears leaving, as a minimum, the internal resistance of the graphite column. One of these resistor elements for the control of power machinery is shown in Fig. 217.

Resistor units of this type present several interesting features. The variation of resistance is stepless, or, more accurately, it occurs through an infinite number of steps. Sliding contacts and the wiring necessary to connect them to the fixed resistance units of the ordinary rheostat are absent. Also the heat radiating capacity of the resistor remains practically constant throughout its range of operation. This is due to the fact that the entire resistor is always employed from one end of the range to the other, whereas in the ordinary type parts of it are successively cut out as the resistance is reduced.

Figure 218 shows a type of carbon-disc resistor adapted to the filament current control of vacuum tubes and similar purposes.

Figure 219 illustrates a carbon-disc rheostat developed for laboratory purposes. It has a resistance range ratio of about 100 to 1. According to the character of the discs chosen, a wide choice of resistance ranges is available. Thus, a range as low as from about 0.01 to 10 ohms or one as high as from about 30 to 3,000 ohms may be had. An interesting and convenient feature of this laboratory type is the intermediate terminal. This is attached to a slotted sheet of copper which may be inserted between the discs at any point, thus providing for intermediate taps as required.

Near the beginning of this chapter it was stated that one of the uses of resistors was to develop heat at desired points. Various electrical heating devices, such as flatirons and soldering irons, are obvious examples of this use. An interesting example

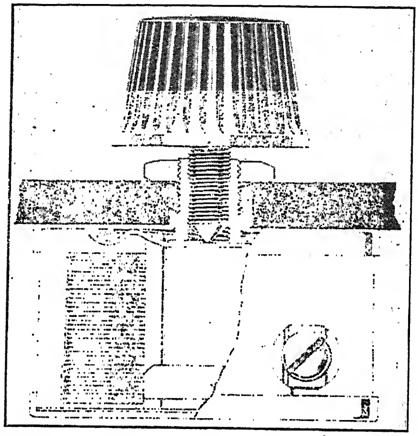


FIG. 218.—Small carbon-disk rheostat. *(Courtesy of Allen-Bradley Co.)*

in telephone work is found in the so-called "heat coil," or "sneak-current arrestor," largely employed in protecting telephone circuits and apparatus against the flow of dangerous abnormal currents. In these a short length of resistance wire is coiled

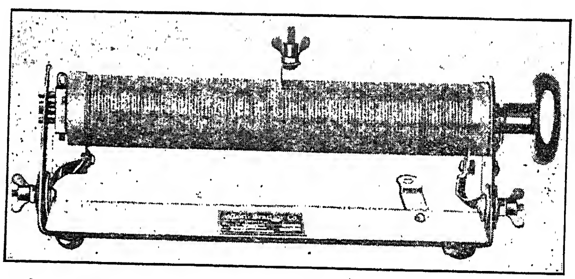


FIG. 219.—Carbon-disk rheostat—laboratory type. *(Courtesy of Allen-Bradley Co.)*

within a space not much larger than a match head. They are placed in series with the apparatus they protect and are usually directly in the paths traversed by voice currents. For this reason they are non-inductively wound so as to present negligible impedance.

Their function is to become heated when traversed by extraneous currents, considered too large for safety, and thus to melt a drop of low-fusing solder immediately adjacent to the core of the coil. The melting of this solder releases a spring which performs certain switching operations intended to safeguard the circuit. These heat coils will be more specifically considered in the chapter on protection.

CHAPTER XIV

CONDENSERS

In broad aspect, any two conducting bodies insulated from each other may be considered as forming an electrical condenser. The amount of electrical charge that the conductors will receive and hold at a given difference of potential will depend on the proximity of the two conductors to each other and on the chlaracter of the insulating material separating them. Two couds floating in the sky, or a single cloud with the earth below, may be considered as a condenser, the phenomena of thunder and lightning being but manifestations of electrical discharges across the air space between its conductors. Similarly, the two wires of a metallic circuit telephone line, together with the insulation between them, exhibit the properties of condensers, and the electrostatic action between the two conductors or between each of them and ground is a factor to be reckoned with in the transmission of telephone currents.

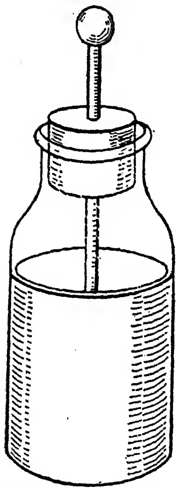


FIG. 220.—Leyden jar.

More specifically, however, the term "condenser" is applied to a piece of electrical equipment comprising two conducting surfaces held closely together but separated by a thin insulating material called a "dielectric." The earliest form of condenser in this more specific sense is the well-known Leyden jar illustrated in Fig. 220. This consists of a glass jar, the inner and outer surfaces of which are coated with tinfoil up to within a few inches of the neck, the inner coating being electrically connected with a metallic ball or other conductor attached to a rod projecting through the neck.

The purpose of the condenser is to enable the conducting plates to receive and hold charges of electricity by virtue of their relatively large area and their close proximity. A positive

charge on one plate attracts and holds a negative charge on the other. The amount of electricity which the two plates are thus able to hold at a given difference of potential varies inversely as the distance between the plates, directly with the area of the plates and directly with an inherent characteristic of the insulating medium known as its "specific inductive capacity."

The measure of a condenser with respect to the amount of electricity it will hold at a given potential is termed its "capacity" or its "capacitance." The capacitance of a condenser, therefore, depends on three different characteristics: first, the area of the plates presented to each other, or, more properly, the area of the dielectric lying directly between them; second, the specific inductive capacity of the dielectric; and, third, the thickness of the dielectric. Obviously, therefore, if we wish to secure a condenser of large capacity in a small space, we choose thin plates of as large area as possible and separate them by a dielectric of high specific inductive capacity as thin as mechanical and electrical conditions will allow.

In the choice of a dielectric, however, other requirements than that of high specific inductive capacity must be met and, in fact, are even more important. Electrically the dielectric material must have sufficient "dielectric strength" to prevent its breaking down under the potential to which it is to be subjected. Also it must have high specific resistance to minimize electrical leakage or "leakance" between the plates. Mechanically it must have the necessary strength and continuity of structure to keep the plates from coming into contact with each other. Chemically it must be stable enough not to change its composition and thus its electrical characteristics with the lapse of time. It should be non-hygroscopic as far as possible, to avoid absorbing moisture with consequent lowering of its insulation resistance.

Considering insulation alone, the American Telephone and Telegraph Company gives the following list of materials in the order of their insulating properties:

Dry air.
Shellac.
Paraffin.
Paraffin paper.
Paraffin oil.
Ebonite.

Rubber.
Porcelain.
Sulphur.
Glass.
Mica
Silk.
Varnish.
Dry paper.
Celluloid.
Dry wood.
Slate.
Fiber.
Distilled water.
Alcohol.

All of these except the last two are used for insulating materials in telephony, and many of them, singly or in combination, as the dielectric materials of condensers. It is interesting to note that water, when in a chemically pure state, ranks as an insulator.

In Table XIV are shown the dielectric strengths and the specific inductive capacities of a number of insulating materials. The wide range given for some of the materials is due largely to the variations often found in the characteristics of different materials bearing the same name. In some cases, also, the previous history of the material will affect markedly its electrical characteristics. Congealed paraffin, for instance, will show con-

TABLE XIV.—ELECTRICAL PROPERTIES OF DIELECTRIC MATERIALS

Material	Dielectric strength, kv. per mm.	Specific inductive capacity
Air	....	1.
Asphalt	14.6	2.7
Bakelite, C-1	Up to 27.5	5.2 to 9.9
Bakelite, wood molding mixture	17.7 to 21.6	4.5 to 5.5
Bakelite micarta, 213	Up to 31.4	5
Bakelite, dilecto	25.6	5
Vulcanized fiber	3 to 16.7	5
Glass	8 to 9	5.5 to 9.1
Mica	21 to 28	5 to 7
Paper	8.7	2.6
Paraffin	11.5	2.1
Rubber, hard	70.	2 to 3.5

siderably lower specific inductive capacity if it has been cooled rapidly than if cooled slowly. As in the case of metals used for conductors and for magnetic purposes, impurities in dielectrics affect their insulating qualities. Such a compilation as that of Table XIV must, therefore, be used only as a general guide in the absence of determinations for each specific grade of a given material.

For reasons that are fairly obvious, mica, paper saturated in paraffin, glass and air are the principal materials used for dielectrics in condensers. Mica, because of its fine insulating character, its high specific inductive capacity, its chemical stability and the facility with which it may be split into thin sheets, comes very near to being the ideal material. Its cost, however, is high, both with respect to the material itself and its fabrication into suitable sheets. In spite of its cost, mica is largely used in the finer grades of condensers where great exactness of capacitance and permanence are prime requisites.

The great majority of condensers used in telephone work employ thin bond paper saturated with paraffin or some similar insulating compound. The reasons for the large use of paper are its cheapness, the ease with which it may be handled in thin sheets, and the fact that when properly treated its electrical and mechanical characteristics remain fairly constant.

For small capacities, and particularly where it is desirable to adjust the capacity between maximum and minimum values, condensers employing air dielectrics are largely used. An adjustable one of these "air condensers" is shown in Fig. 221. In this there are two sets of interleaved plates, the plates of each set being electrically connected to each other but carefully insulated from those of the other set. One set of plates is fixed while the other is capable of rotation on a shaft turned by a hand knob. The position of maximum capacity is, of course, that in which the movable plates lie entirely between the stationary ones, while minimum capacity is secured when the movable plates are withdrawn as far as possible from between the stationary ones. The maximum capacity of this particular condenser is about 0.0015 microfarad. Sometimes such condensers are immersed in oil. Air or oil condensers of this sort are much used in radio work.

Where great exactness and constancy of capacitance together with reasonable compactness are paramount requirements,

mica is the most used dielectric for condensers. In forming the mica condenser alternate sheets of tinfoil and mica are laid up in a stack. The mica extends beyond each sheet of tinfoil on three sides to avoid contact between adjacent conducting plates. Alternate sheets of tinfoil, however, extend beyond the mica sheets on opposite sides of the stack, and, when the stack is completed, the sheets projecting from each of its two

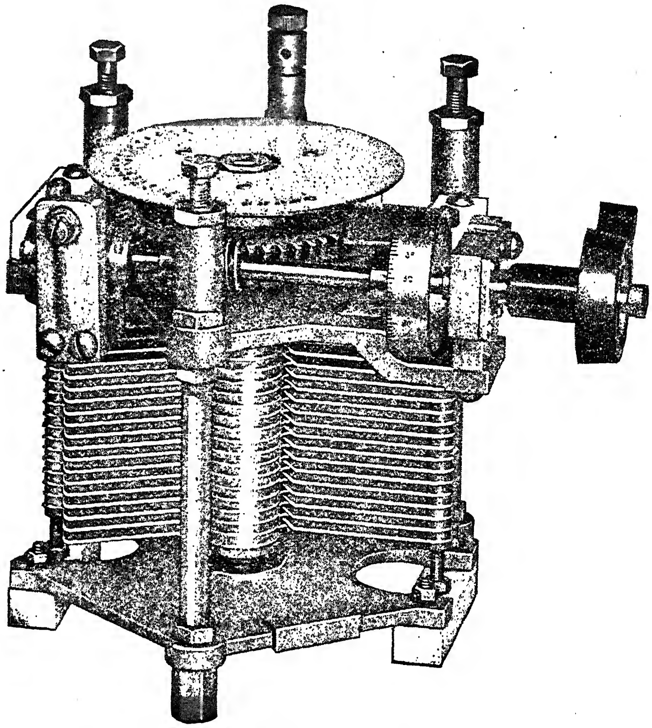


Fig. 221.—Adjustable air condenser. (*Courtesy of John Wiley & Sons, Inc.*)

opposite sides are connected together. The whole is then compressed, put into a metal receptacle and filled with a suitable insulating compound, which, on hardening, holds the entire mass in permanent relation and effectively seals it against the entrance of moisture. The capacities of such condensers, properly made, remain nearly constant for long periods of time, varying only slightly with temperature. When properly calibrated for different temperatures, their capacitance for any given temperature may be relied on to an accuracy of perhaps one part

in ten thousand. Such extreme accuracy as this is required only for standards in making precise measurements. There are, however, uses in commercial telephone work where both accuracy and constancy are required. An instance is in the condensers employed in the filters of carrier-current systems. Here it is required that the variations in capacitance due to such temperature changes as occur in a central office and due to aging over long periods of time shall be within 1 per cent.

To obtain greater precision in capacity than is ordinarily afforded by the method of stacking alternate sheets of mica and tinfoil, sheets of brassfoil are used for a few of the conducting

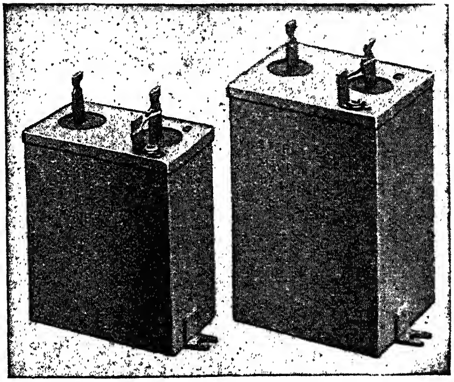


FIG. 222.—Mica condensers of high accuracy and constancy. (*Courtesy of Bell Telephone Laboratories.*)

sheets, and the condenser is built up to slightly greater capacitance than that desired. Then, after the condenser is tested, the exact capacitance is obtained by pulling some of the brass sheets partly out from the pack. The reason for using brass instead of tinfoil for the removable sheets is its greater strength in resisting tearing.[1] Two such dry-stacked condensers of high accuracy and constancy as manufactured by the Western Electric Company are shown in Fig. 222.

For most uses of telephony no such accuracy or constancy as that possible with the mica condenser are required. Variations in capacitance as great as plus 25 per cent from the rated value are often permissible. In some cases, however, the requirements demand that the capacities be held within rather narrow limits

[1] YOUNG, C. R., Condensers for Many Uses, *Bell Laboratories Record*, June, 1929.

and that they remain quite constant. This is particularly true in those electrical circuits and networks where resistances, inductances and capacitances must bear close relationship with respect to each other and to the range of frequencies with which they are to be used. Again, it is often desirable that condensers be closely matched in pairs to secure a proper balance of the circuits in which they are used. This is true, for example, of condensers used in the simultaneous sending of telephone and telegraph messages over the same circuit.

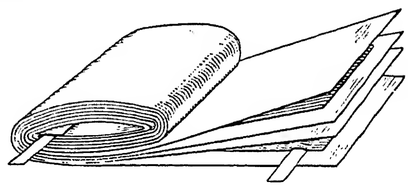


FIG. 223.—Method of making rolled-paper condensers.

For all ordinary purposes of telephony, except where very great accuracy and constancy are required, the paper-insulated condenser is admirably adapted, and its cheapness makes for its almost universal use. At first these were made by laying up alternate sheets of tinfoil and oiled paper after the manner used in making mica condensers. The requirement for large quantity production, however, soon led to a very much more economical method, resulting in the so-called "rolled condenser."

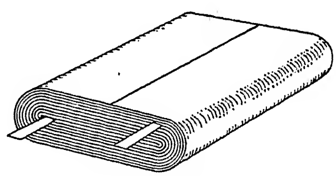


FIG. 224.—Condenser roll before encasing.

In the rolled condenser the two conducting plates and the intervening paper are rolled up in long continuous strips instead of being formed of many small sheets piled on top of each other. The manner of construction is indicated in Figs. 223 and 224. The paper, a fine, thin grade of bond of uniform texture and as free as possible from metallic or other conducting particles, is purchased in rolls of uniform width. The tinfoil that is to form the plates is also purchased in rolls, very thin and of slightly less

width than the rolls of paper. Two rolls of tinfoil and four of paper are then wound on a flat mandrel, in such a way that two layers of paper always lie between the adjacent surfaces of the tinfoil. Two thin metal terminals are wound into the structure, the first making contact with the inner end of one sheet of tinfoil and the second with the outer end of the other. These project beyond the edges of the paper and afford convenient means of connecting with the external terminals.

The employment of two or more very thin strips of paper, as in Fig. 223, instead of a single, thicker one, safeguards the condenser against the short-circuiting which might result from pinholes or small metallic particles if only a single sheet were used. With two sheets, even if pinholes or metallic particles do occasionally occur, the chance of their lying exactly opposite each other is remote, and their harmful effects are almost negligible.

Another form of rolled condenser known as the "Mansbridge," was introduced in London about 1908. This employs a metal-coated paper instead of separate sheets of tinfoil and paper as in the condenser just described. In rolling the Mansbridge condenser, two sheets of metal-coated paper with alternate sheets of thin, uncoated paper are rolled together in a manner similar to that just described. The Mansbridge type of condenser is now little used in this country—the type employing separate sheets of tinfoil, separated by two sheets of paper (Fig. 223) being almost universally employed.

After rolling, the unit is baked to expel all moisture, boiled in paraffin, and then subjected to slight pressure on the flat sides and allowed to cool under this pressure. The rolls so formed (Fig. 224) are complete condensers in themselves, but require mechanical protection to render them sufficiently hardy to stand handling. They also require treatment to assure their retaining as far as possible their original form, size, capacity and insulation resistance.

For these reasons, after the paraffin-saturated rolls have cooled, they are given a quick dip in a molten insulating compound which hardens at a temperature considerably above ordinary atmospheric temperature. A compound that has been much used by the Western Electric Company for this purpose is composed of about 80 per cent asphaltic cement, 8 per cent rosin and 12 per cent Montan wax. Such a compound melts at about 400° F., and is sufficiently hard and tough at atmospheric

temperatures to afford the desired rigidity and strength. This dipping gives the roll an external layer of this compound. After again cooling, the rolls are inserted in sheet-metal cans of the desired size and shape, and the remaining space then filled with the same asphaltic compound.

Condensers so treated are sufficiently permanent in capacity and insulation resistance for most uses. It is found, however, that they do undergo a slow change in their electrical characteristics, due probably to a gradual change in the relationship between the plates caused by a redistribution of internal stresses with time and changing temperatures. To give still greater permanency, for the more exacting requirements, the finished condensers are sometimes placed in an oven which is maintained at a temperature somewhat greater than the melting point of the impregnating wax, held at this temperature for several hours and then allowed slowly to cool. This permits the internal stresses to become equalized while the whole mass is taking its final set. Condensers so treated show little tendency to permanent change under normal operating conditions.

It is not possible, or at least not commercially practicable, to make rolled-paper condensers which, when finished, will have exactly a predetermined required capacity. Even with the greatest care to preserve uniformity, there will be some variation in the capacities of condensers in the same batch. Where greater exactness of capacity is required than can be secured in manufacture, it is attained by a process of selection.

The condensers of a given batch having approximately the desired capacities are individually measured, and those having the exact capacities desired for a particular purpose are chosen. Those whose capacities fall outside the desired limits are used for other less exacting purposes.

Rolled-paper condensers possess sufficient dielectric strength to withstand the maximum voltages normally employed in telephone work. They are ordinarily submitted to and able to withstand a breakdown test of 500 volts. This is far above the voltages employed in telephone work, and considerably above those from outside sources which the protective appliances are supposed to guard against.

The ordinary range of capacity for single rolled condensers is from a few hundred-thousandths to two microfarads. Figure 225 shows two examples. The larger one is the No. 21-B W condenser

of one microfarad capacity used in the ringing circuit of Western Electric common-battery telephones. This general type uses 5-mil paper, 4 inches wide and requires about 18 feet for each microfarad of capacity.

The smaller one of Fig. 225 is the No. 129-A condenser used in the transmitter case of the new Western Electric hand telephone set. This has a capacity of about one-hundredth of a microfarad. Larger capacity units than two microfarads are sometimes made commercially, but usually, when large capacities are desired, they are obtained by connecting as many units in multiple as may be required. As will be shown, the combined

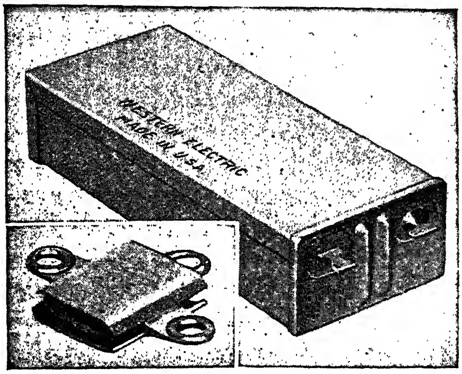


FIG. 225.—Standard condensers for use in telephone sets. (*Courtesy of Bell Telephone Laboratories.*)

capacity of condensers connected in multiple is the sum of their separate capacities.

The size and shape of commercial rolled condensers depends on the standards adopted by the manufacturing company as well as on the capacity and the available space in which they are to be mounted. For purposes of economy and standardization, manufacturers try to adopt as few sizes of can as possible, and to employ each of these for condensers of several capacity ratings even though a considerable amount of waste space in a can is involved. To illustrate: The Stromberg-Carlson Company employs exactly the same size can, $4\frac{3}{4}$ by $1\frac{7}{8}$ by 1 inch, for capacities of 1, 2 and $3\frac{1}{2}$ microfarads in one of its condenser styles.

Condenser terminals are usually of the rigid type, adapted for soldering the connecting wires. Sometimes, however, flexible terminals are employed, and in other cases the well-known

Fahenstock clip is used. Several shapes of commercial rolled condensers with different types of terminals, as made by the Stromberg-Carlson Co., are shown in Fig. 226.

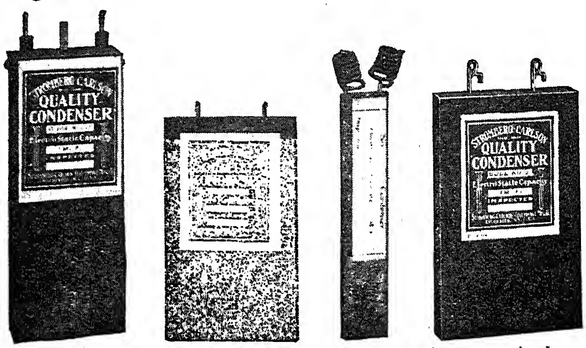


Fig. 226.—Commercial condensers with various terminals.

The mounting of condensers in convenient relationship with respect to other elements of apparatus with which they are to be associated assumes a great variety of forms. When forming parts of telephone sets, they are usually strapped in position by

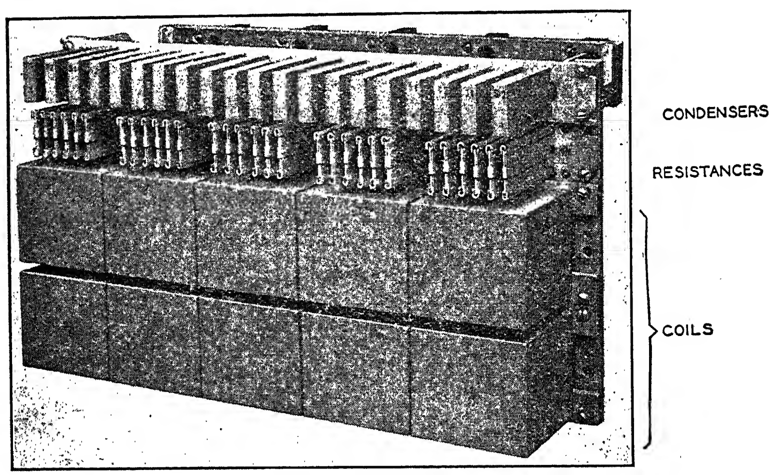


Fig. 227.—Condensers on common mounting plate. (*Courtesy of Bell Telephone Laboratories.*)

a simple metal clamp. In connection with central-office equipment the same practice may be followed for individually mounted units, but, where they are used in groups, it is common practice to support them on steel mounting plates conforming in general

dimensions and arrangement with the mounting plates carrying other parts of switchboard apparatus, such as coils, resistors and relays. Figures 227 and 228 show such multiple mountings as practiced by the Western Electric Company.

Frequently, the requirements of close mounting are met in switchboard work by sealing a number of small condensers in a single can, the terminals of each individual condenser being brought out separately. In carrying out this idea the Western

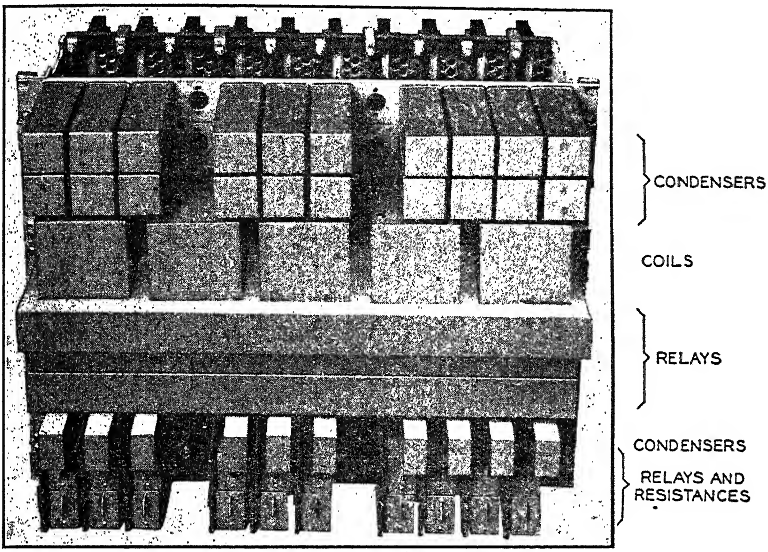


FIG. 228.—Group mountings of switchboard equipment. *(Courtesy of Bell Telephone Laboratories.)*

Electric Company makes one unit which contains 17 separate condensers, one for each of the 17 cord circuits of a switchboard position.

Where paper condensers are required to withstand higher voltages than those ordinarily encountered in telephony, more layers of paper than two are interposed between the conducting plates. A 20,000-volt condenser of large capacity is shown in Fig. 229. This employs 5 sheets of thin paper between the strips of tinfoil. This condenser is used in the filter power circuit in the transatlantic radio transmitting station at Whippany.

The recently developed cellulose material, known as "cellophane," which comes in very thin transparent sheets and is used

as a covering for boxes of candy and other perishable commodities, has attracted attention as a possible substitute for paper as a condenser dielectric. Unfortunately, however, in order to give this material its flexibility, glycerine is used as a softening agent in its manufacture. This is objectionable because glycerine is hygroscopic, attracting and holding moisture

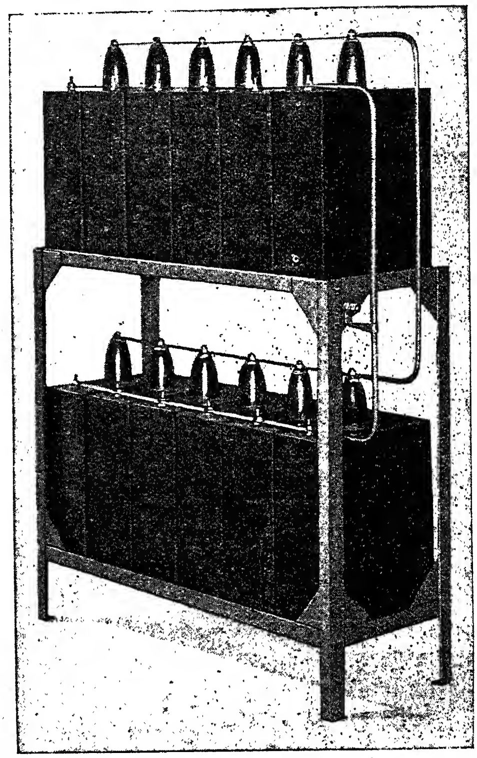


FIG. 229.—20,000-volt paper condenser. (*Courtesy of Bell Telephone Laboratories.*)

readily. When manufactured without this softening agent, sheets having higher insulation resistance and breakdown strength than any of the condenser tissues now used were produced. Their extreme thinness made possible high capacities in small volume. The difficulty with this, however, was that it was brittle. After being wound, dried, impregnated with wax and pressed, as in the rolled-paper condenser, cracks developed at the curved end, resulting in high leakage. It is possible that

further development work will overcome these difficulties, and that a more nearly ideal cheap condenser dielectric will result.

We come now to a comparatively recent development, that of the electrolytic condenser—a radically different type from those already considered.[1] This has only recently assumed an important place in telephone work, strangely enough in connection with power plants where condensers have heretofore been little used.

The electrolytic condenser has physical characteristics more like a storage battery than a condenser, but it possesses, nevertheless, the fundamental attributes of the orthodox condenser—two conducting surfaces separated by a thin insulating dielectric.

When two plates of aluminum are immersed in a suitable electrolyte and subjected to a continuing difference of potential which tends to make current flow between them, the anode or plate from which the current flows will, after a time, develop an insulating film on its surface. The film slowly disappears, as though dissolved, if the potential is withdrawn, but will remain indefinitely if the potential is continuously maintained. In this condition we have, then, the remarkable phenomenon of two metal plates immersed in a conducting fluid yet completely insulated from each other.

The nature of the film formed on the anode plate and insulating it from the electrolyte is not fully understood. It has high insulating qualities, however, and is extremely thin. The actual thickness of the film on aluminum plates has been determined to be from 0.001 to 0.0001 millimeter, depending on the voltage and other conditions of formation. On account of its thinness and its high insulation resistance, it forms an extremely efficient condenser dielectric between the two condenser surfaces, one of which is the positive aluminum plate and the other the adjacent surface of the electrolyte. Evidently, the film becomes thicker as the applied potential is raised, for, as will be shown, the capacity of a condenser so formed decreases as the applied voltage is raised.

The construction of the aluminum plates for such a condenser is shown in Fig. 230. The positive plates are corrugated to afford large surface; the alternate negative plates are flat, their

[1] The Aluminum Electrolytic Condenser, by H. O. Siegmund, presented before American Electrochemical Society, Bridgeport, Connecticut, April 26, 1928; also in *Bell System Technical Journal*, p. 41, January, 1929.

function being merely to make contact with the electrolyte, the surface of which, resting against the insulating film, forms the other condenser plate. The plates are bolted to a porcelain cover which holds them in proper space relation.

The electrolyte is a mixture of ammonia, boric acid and water. It is held in a heavy glass jar in which the assembled plates are immersed. After assembly, the electrolyte is covered by a layer of paraffin oil to prevent evaporation. The appearance of the completed cell is shown in Fig. 231.

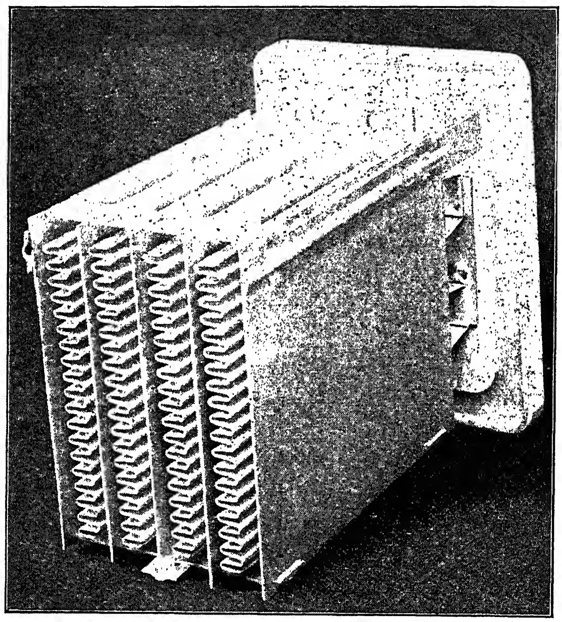


FIG. 230.—Construction of electrolytic condenser. (*Courtesy of Bell Telephone Laboratories.*)

This cell, as developed by the Western Electric Company for telephone power-plant use, is 8 inches wide, 10¼ inches long and 14¼ inches high. Completely assembled, with electrolyte, it weighs 42 pounds. On a 24-volt direct-current charging circuit it has a normal capacity rating of 1,000 microfarads at 1,000 cycles. On a 48-volt charging circuit its capacity is about 600 microfarads at the same frequency.

Such a condenser, so far as experience has shown, requires practically no maintenance. The solution and the plates have shown their ability to last for several years without renewal

or apparent deterioration. The leakage current through them is negligible.

Up to this time electrolytic condensers are finding their principal application in telephone power plants, where very large capacities are now being used across the charging mains to strain out the noise-producing ripples introduced into the charging current by the charging generators and signaling apparatus. It remains to be seen what may be developed from them where smaller capacities are required. The fact that they contain a corrosive liquid obviously renders their use objectionable in many places for which the paper condenser is admirably adapted. On the other hand, the securing of 1 microfarad capacity with plates having surfaces about as large as a postage stamp is alluring, and it is to be expected that the electrolytic condenser will find other fields of usefulness in the future development of telephony.

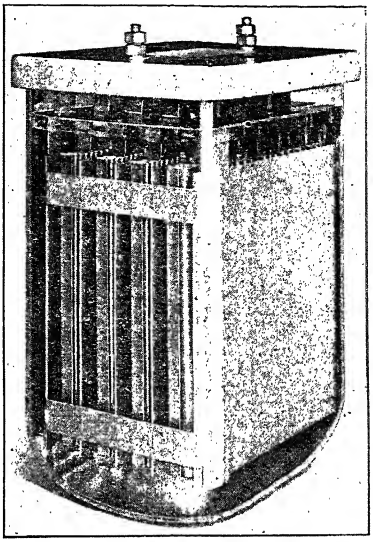


Fig. 231.—Complete electrolytic condenser cell. (*Courtesy of Bell Telephone Laboratories.*)

Speaking again of condensers generally, the functions they perform in the telephone art are of four principal kinds:

1. They present a path of such nature as to prevent the flow of direct currents and to allow the flow of alternating currents. This is exemplified by the use of a condenser to afford a by-pass for voice currents around the impedance of a relay coil that must, for other purposes, be included in the circuit.

2. They act in connection with resistances and reactances in such manner as to present paths which offer little obstruction to certain frequencies and great obstruction to others. This point is exemplified in tuned circuits and in filters that act to pass certain bands of frequencies and exclude others. This point is really a refinement of the first.

3. They act to absorb and give back charges of electricity in proper time relation to current variations in associated cir-

cuits, so as to reenforce or "boost" these variations. An example of this is in the so-called "booster circuit" of the standard common-battery telephone instrument employed by the Bell companies.

4. They act to reduce or suppress destructive sparks that might otherwise occur on the sudden breaking of an inductive circuit carrying a current. When a reactive circuit carrying current is suddenly broken, a spark usually results at the break, due to the extra current or "kick" of the reactance. A condenser bridged about the break may be made to absorb this momentary extra current, thus reducing or obliterating the spark and avoiding its destructive effects.

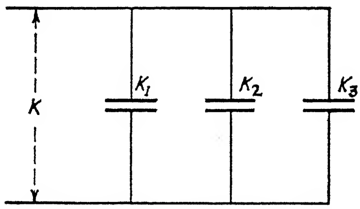


FIG. 232.—Condensers in multiple.

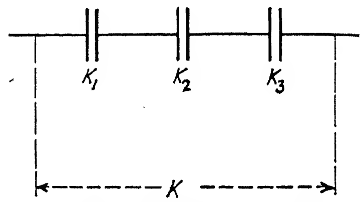


FIG. 233.—Condensers in series.

When two or more condensers are connected together in parallel relation (Fig. 232), their combined capacity is equal to the sum of their individual capacities. Thus,

$$K = K_1 + K_2 + K_3 + \cdots$$

If, however, they are connected in series, as in Fig. 233, their combined capacity will be less than the capacity of either of them alone. The mathematical relationship in this case may be most easily remembered by saying that the reciprocal of the combined capacity will be the sum of the reciprocals of the individual capacities. Thus,

$$\frac{1}{K} = \frac{1}{K_1} + \frac{1}{K_2} + \frac{1}{K_3} + \cdots$$

It was shown in Chap. VIII that in an alternating-current cirucuit containing only capacitance, the current leads the electromotive force by an angle of just 90 degrees. In such a circuit, if it could exist, the power consumed would be zero. Electricity would merely surge back and forth, charging and discharging the condenser without loss. The current, though perhaps large, would be truly wattless. That this is so may be seen from

the power equation, $P = EI \cos \theta$. Since θ, the angle of lead, is 90 degrees, its cosine, which is the power factor, is zero. The power is then zero, regardless of any values E and I may have.

Such a circuit would require not only that its conductors possessed no resistance but also that the condenser be a perfect one, with no power losses of any kind.

Practically, a perfect condenser is an impossibility, although a close approach to perfection may be attained. Always some losses occur and, for ordinary voltages and frequencies, they are due to three different causes: leakage, or leakance as it is frequently called, dielectric "absorption" and resistance of the terminals and plates.

The leakage loss is due to the actual leakage of current from one plate to the other through the dielectric or around its edges. Even though the specific resistance of the dielectric may be very high, the path through it is short and its cross-section large, so the actual flow of current through it may not be of negligible amount. The amount of leakage for any given voltage is dependent on the insulation resistance of the condenser, that is, the resistance of the path across the dielectric. To give a quantitative idea of this, in one instance, it may be stated that it is a common requirement in the manufacture of rolled-paper condensers of 1 microfarad capacity that they shall have an insulation resistance of not less than 500 megohms, Whatever loss of energy occurs by leakage is, of course, an ordinary I^2R loss the energy being wasted in heating the condenser.

The absorption loss in condensers is a loss occurring within the dielectric itself. It is none too well understood. It has been called "dielectric hysteresis" on the theory that it was akin to magnetic hysteresis in iron and caused by some sort of molecular rearrangement within the dielectric under the varying electrostatic stresses. The more modern view seems to be that there is no exact parallel to magnetic hysteresis and that dielectric absorption is an I^2R loss, which makes it difficult to distinguish from leakage loss.

The remaining kind of condenser loss, that due to the ohmic resistance of the terminals and leads and of the plates themselves, is a pure I^2R loss through the conductors of the condenser. It is, therefore, spent in heat. It is, naturally, more pronounced in those condensers whose plates consist of two long strips of tinfoil in continuous rolls than in the stacked types in which the

plates are built up of many single sheets directly connected together at their edges.

The ideal condenser, with no internal losses, would be one with infinite insulation resistance, with zero absorption effect and with zero resistance in its terminals, leads and plates. In so far as these ideal conditions are unattainable in practice, the losses just described will exist under alternating or varying electromotive forces. Since all of these losses are heat losses, their combined effect is the same as if a resistance of such value as to produce an equal I^2R loss were placed in series with a theoretically perfect condenser of the same capacity.

The part of the total current that is involved in these losses, not being wattless, is in phase with the electromotive force. It thus reduces the angle of lead to a value somewhat below the theoretical 90 degrees. The power factor of a condenser, then, is not zero, as it would be for a perfect condenser, but is equal to the cosine of an angle somewhat less than 90 degrees, by which the resultant current leads the applied electromotive force.

CHAPTER XV

FLEXIBLE CORDS

Flexible cords, as the term is used in electrical work, are insulated electrical conductors constructed with special regard to their flexibility. They may contain one or several conductors. They are used to connect parts that must be conductively related but relatively movable. Their principal use is for permitting a certain degree of portability to parts that must be electrically connected to other parts that are stationary.

Familiar examples are ordinary duplex lamp cords for lending portability to incandescent lamps; receiver cords connecting telephone receivers with the other parts of the telephone set; desk-stand cords connecting portable desk telephones with the stationary parts of the house wiring; and switchboard cords by which the operators at manual switchboards establish connections with telephone lines.

Conductor flexibility is attained either by spiralling the conductor or by building it up of many smaller conductors, or both. Where the spiralling method alone is employed, such great flexibility as that of a closely coiled spiral of wire may be attained, but at a loss of tensile strength and, of course, with an increase of conductor resistance. In such a case the required conductivity and tensile strength must be secured by supplemental means.

In general, efforts to attain a maximum of flexibility are likely to result in a loss of tensile strength or of conductivity. These qualities, however, must be retained in the required degree if the cord is successfully to serve its purpose. Also, durability is of prime importance not only with respect to wear and tear in the mechanical sense but also with reference to maintaining the required conductivity and insulation.

Where considerable current-carrying capacity is a factor and extreme flexibility is not required, the conductors may be made up of small copper wires twisted together in a single strand. The conductors of lamp cord are formed in this way from No. 30

B. & S. gage wires in sufficient number to give the required conductivity. The stranded bare conductor is covered successively with a serving of cotton, a wall of rubber compound and, finally, with a braid of cotton or silk. Two such insulated conductors twisted together form ordinary duplex lamp cord.

Somewhat greater flexibility than that of a single-stranded conductor may be secured by twisting together several multiwire strands into the form of a rope. Another method of making flexible conductors is to run the small copper wires either singly or in groups through a braiding machine, resulting in a tubular braided strand of the general appearance shown in Fig. 234. Such flexible braided conductors may vary in size from the very small conductors used to form "pigtail" terminals for electromagnets up to large conductors having an equivalent size of several hundred thousand circular mils.

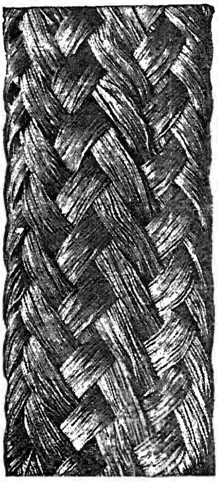


FIG. 234.—Braided conductor. (*Courtesy of Belden Manufacturing Co.*)

Flexible conductors made by variously combining small round wires by any of the foregoing methods are only of incidental use in telephone work. Here the requirements are often different from those in the power and lighting field. Current-carrying capacity is usually relatively unimportant, while flexibility, durability, constancy of conductivity and certainty of insulation are paramount.

Very thin ribbons of copper or bronze, called "tinsel," play an important part in the manufacture of telephone cordage. The threads of tinsel are sometimes stranded or braided or formed into a rope to provide a conductor of great flexibility. More often, however, as will be shown, the tinsel threads are wrapped spirally around individual threads of some textile material. In this way the conducting thread is given added flexibility, while all tensile strength is supplied by the textile.

Coming now to specific instances of telephone-cord construction, a switchboard cord that depends for its flexibility entirely on the spiralling of the conductors is shown in Fig. 235. In this a conductor of flexible steel wire is given a smooth close wrap of copper wire to increase its conductivity and is then insulated

with a tussah silk serving and a braid of waxed cotton. Two conductors, so insulated, are coiled into a double helix. A braid of soft cotton and an outer braid of glazed cotton are then placed over all.

The plug end of this cord has an extra braid extending back about eighteen inches to give additional wearing qualities at that end. To further reinforce the plug end and to guard

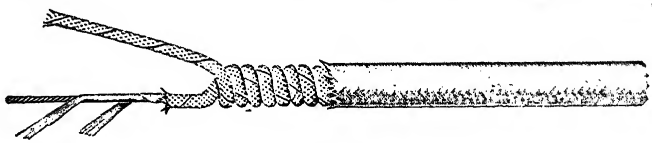


Fig. 235.—Spiral-type cord.

against its being kinked too sharply, a short length of steel wire covered with a silk braid is used as a core extending a short distance back from the plug.

Another cord made on an entirely different principle is shown in Fig. 236. In this a conductor of 18 threads of copper tinsel is twisted into a 3-cord rope. This is covered with two wrappings of tussah silk, the outer one of which is saturated with a moisture-proof compound. A dry cotton braid forms the third layer of insulation on each conductor. One of the three con-

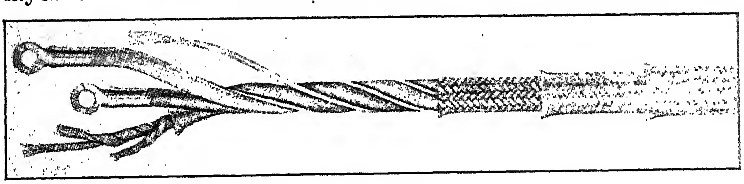


Fig. 236.—Tinsel cord.

ductor ends in this cut is shown dissected, exposing successively the tinsel rope, the two silk warppings and the cotton braid. The other two conductors are shown with their terminals attached, as in a finished cord.

After laying up the requisite number of such insulated conductors in cable form, with textile fillers to make it round, they are securely bound together by a braid of cotton. About eighteen inches from the plug end another braiding of glazed cotton is started and continued toward the plug end, after reaching which it is continued back over the entire length of the cord. This gives a reenforcing braid at the plug end of the cord where the wear is most severe, resulting in three braids at the

plug end and two throughout the remaining portion of the length. This reenforcement is common practice in the manufacture of all modern switchboard cords.

FIG. 237.—Old-type spiral-steel cord.

An older type of spiral wire cord than that of Fig. 235 is shown in Fig. 237. This has for its center a core of strong lock-stitch twine, which is primiarly depended upon for tensile strength. On this central twine is woven a braiding of tinsel and on this is wrapped a helical steel wire, which, together with the tinsel braiding, forms the inner conductor of the cord. Over this inner conductor are then applied, successively, a braiding of tussah silk, a braiding of linen and another braiding of tinsel. On this second braiding of tinsel the second steel wire helix is wound, this being covered by one or more braidings of linen or glazed cotton subsequently saturated with some moisture-resisting compound.

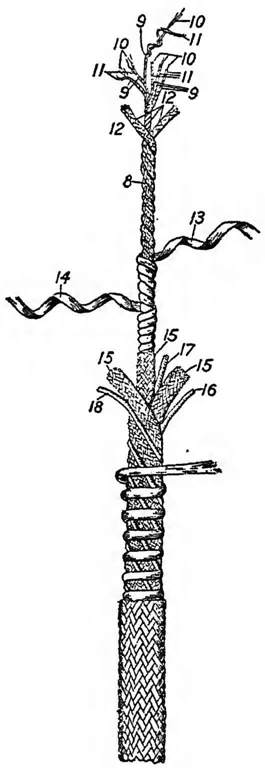


FIG. 238.—Recent Western Electric tinsel cord.

In this cord the steel spiral conductors are principally relied upon for continuity of the conducting paths, while the presence of the layers of copper tinsel on which the spirals are wrapped reenforces the conductivity.

An objection of this type of spiral steel, especially under exacting voice transmission requirements, is the comparatively high inductive capacity between the two talking conductors, caused by their large concentric cylindrical surfaces lying so close together.

For a number of years and until very recently, the standard cords made by the Western Electric Company, and universally used by the associated Bell companies, have been made in accord-

ance with the principles illustrated in Fig. 238. This shows the method of construction as applied to a three-conductor switchboard cord, but receiver cords, desk-stand cords and others follow the same general plan of construction.[1]

Each of the three conductors, of which one is shown at 8, is made up as follows: A thin ribbon of tinsel, 11, is wound around a textile thread, 10, forming what is called a "twist" or a "tinsel thread," 9. Six of these twists are laid together in a close spiral resulting in a single strand, 12. Three of these strands are then laid together in a rope, forming one of the conductors, 8, of the cord.

This conductor is then wrapped with two servings of tussah silk, 13 and 14, wound tightly in opposite directions. These are subjected to a moisture-proofing treatment after which a braiding of cotton, 15, is applied. In the case of a three-conductor cord, three of these insulated conductors are twisted together in the form of a rope, no core being used. Filler threads of cotton, 16, 17 and 18, are added to make possible a smooth, round exterior finish. The whole structure is then bound together by a spaced serving of cotton and the outer protective braid, usually of cotton, applied.

The moisture-proofing treatment given to the individual conductors of this cord is applied alike to switchboard, desk-stand, receiver and other cords used by the Bell companies. Although such cords are always used indoors and, therefore, supposedly in dry places, the happenings which may expose them to moisture are in the aggregate of frequent occurrence. Among these may be mentioned exposure to live steam or water from leaky pipes or radiators, to rain from open windows or leaky roofs and to spilled ink or other liquids. In the case of switchboard cords which are subject to constant handling, perspiration from the hands of operators is another cause which, unless guarded against, ultimately results in deterioration. It is desirable that moisture-proofing treatment of cords shall not extend to the outer covering, as this would mar the appearance of the cord and interfere with its handling. Also, it should not materially increase the diameter of the cord nor impair its flexibility. In addition, it should not cause deterioration of any of the cord materials.

After the single conductor (Fig. 238) has been covered with its two wrappings of silk, 13 and 14, it is drawn through a hot

[1] U. S. Patent 1,116,090, Nov. 3, 1914.

bath of insulating material composed chiefly of asphaltum. The temperature of the bath and the speed at which the conductor moves through it are such that the outer layer of insulation is completely, and the inner layer only partially, impregnated. The compound reaches the actual conductor scarcely at all. This is done in continuous lengths before the individual braiding, 15, is applied, the conductor passing from one reel through the bath, then through an air blast, if necessary to cool and dry it, and on to a second reel. In this way the conductor is made moisture proof without detracting from its flexibility. Neither the braiding on the individual conductors nor the outer braiding covering the assembled multiconductor cord are impregnated, leaving their appearance and handling qualities unimpaired.

Millions of these cords have been made annually during recent years. Switchboard cords of this type have an average life in service of about two years and require repairs on an average of about once in eight months. In comparison with these figures the old tinsel cord used by the Bell companies up to about 1902 had a life of only about four months and required repairs once a month. The later cord, therefore, lasts about six times as long as the old one and requires repairs about one-eighth as often. Between these two types of cords one having a spiral-steel conductor was largely used. While this gave about the same life results as the later tinsel cord, it possessed the disadvantage, already mentioned, of poorer transmission qualities and, in addition to that, was somewhat stiffer.

A modification of this highly successful method of cord construction has recently become standard practice of the Bell companies.[1] The make-up of a single conductor of this latest cord is shown in principle in Fig. 239. The feature of this modification consists in providing a composite conductor combining a number of strands (six as shown), each of which, in turn, consists of several metal tapes wound on a textile thread. The individual tapes of each strand are wound in the same direction, the outer ones thereby causing the inner ones to bind closer to the core of that strand. The strands are then twisted together on a central core of textile cord, the twist in this case being in the reverse direction from that of the tapes on the individual strands.

[1] *Bell Laboratories Record,* p. 196, July, 1926.

The composite conductor so produced is of great flexibility and not likely to twist or become kinked. By this construction it is possible to use tapes rolled so thin as to permit almost indefinite bending without breaking and yet maintain high conductivity as a result of the tapes overlapping and paralleling each other. In the new standard cord, two metal tapes instead of three are used on each strand. Figures 240 and 241 are enlarged views of single conductors of the recent and the present standards, respectively.

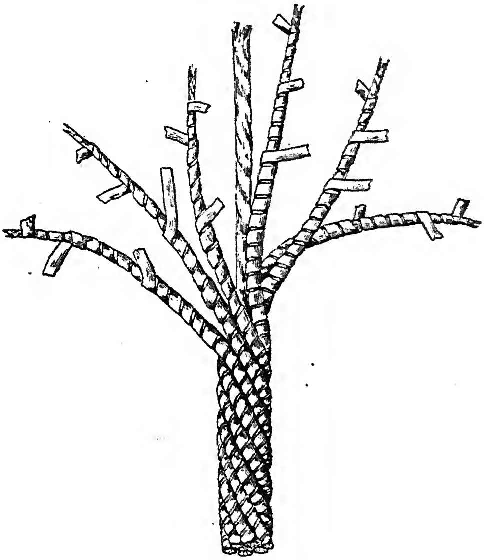


FIG. 239.—Single conductor of new tinsel cord.

A conductor of the recent cord consisted of 18 tinsel threads twisted into a small 3-strand rope, while that of the new cord has only 6 tinsel threads, each made up a of relatively large cotton thread having 2 wider metal ribbons wound around it. The complete conductor of the later type is made by twisting 6 of these double-ribbon tinsel threads around a central core of bare cotton thread.

Information is not available as to the life of switchboard cords constructed in accordance with this new standard, but, according to the article just referred to, it has longer life in service, greater ease in soldering to terminals, greater flexibility and lower cost. The longer life is attributed to the reduced

abrasive action between the tinsel threads when the conductor is flexed. The greater ease in soldering is due to the fact that the two relatively large ribbons more completely cover the cotton threads than the single ribbon of the old tinsel. The increased flexibility is due to the cushioning effect of the bare cotton core and also to the specific way in which the conductors are twisted.

The method of terminating the conductors of finished cords deserves attention. Lamp cord, of course, is ordinarily furnished in long lengths and cut as occasion demands. The conductors in

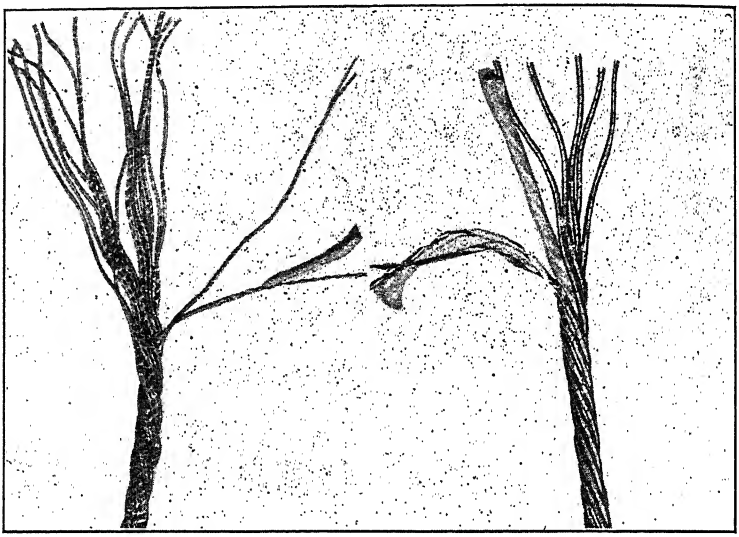


FIG. 240. FIG. 241.

FIG. 240.—Recent Western Electric tinsel conductor.
FIG. 241.—New Western Electric tinsel conductor.
(*Courtesy of Bell Telephone Laboratories.*)

this case are sufficiently hardy to permit their direct attachment under the heads of the binding screws of the lamp socket or of other devices to which the cord is to be attached.

Where cords are manufactured to fit specific pieces of apparatus, such, for instance, as receiver cords and switchboard cords, it is customary to manufacture them in complete units of specified lengths and with terminals attached. The switchboard cord shown in Fig. 242 is an example. The three conductors each terminate at one end in a metallic clip which is soldered to the cord conductor and also made to embrace and clamp the end of the insulating covering to prevent its fraying. The right-hand

end of this cord is shown attached to a switchboard plug. The method of making this connection within the handle of the plug will be described under the discussion of switchboard apparatus.

The exact method of electrically and mechanically attaching cord conductors to their terminals varies greatly with the type of cord. The one shown in Fig. 243 is typical of older practice.

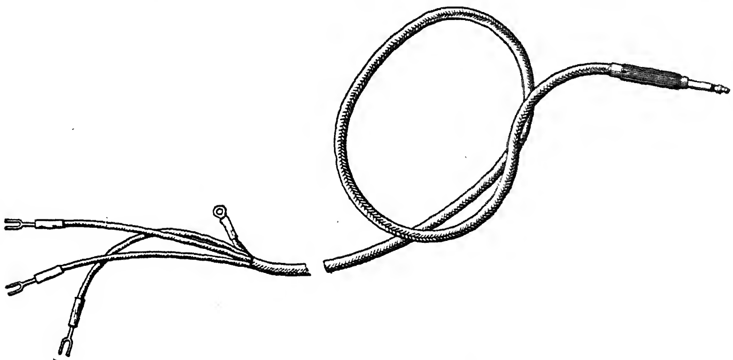


FIG. 242.—Switchboard cord and plug.

Here the projecting tinsel conductor is wrapped with fine copper wire because of the difficulty experienced in making solder stick to tinsel. This wrapping is carried back over the braiding to prevent fraying. With this preliminary preparation, the metal of the terminal is tightly squeezed around the outside of the braid, after which the projecting conductor end is soldered to the terminal.

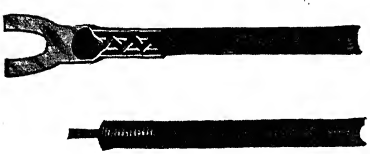


FIG. 243.—Typical cord terminal.

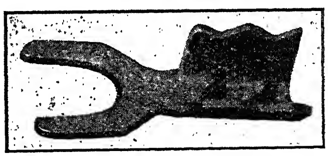


FIG. 244.—New cord tip.

In general, the adequate attachment of terminals to flexible cords has always involved hand operations, more or less tedious and, therefore, somewhat costly. The soldering operation, particularly where tinsel is involved, has always been rather unsatisfactory, on account of the liability of burning the tinsel in case of too much heat, or a failure of the solder to "take" in case of too little heat. Recently, the Western Electric Company has avoided these difficulties by doing away with both the

soldering and the hand wrapping of the conductor ends. The tip shown in Fig. 244 is similar to the older types but has two prongs projecting from the inside of the shank which pierce the insulation and make contact with the tinsel inside. The tips are attached by a machine which presses the prongs through the insulation and then closes the shank of the tip around the cord, thus accomplishing the double purpose of holding the prongs tightly in place and binding the ends of the insulation to prevent fraying. In Fig. 245 are shown the old and the new tips attached to cords, the two right-hand views showing the front and rear of the new tip.

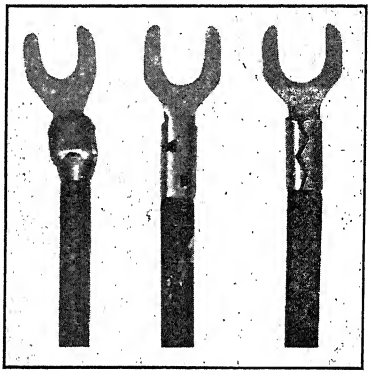


FIG. 245.—New and old cord terminals. (*Courtesy of Bell Telephone Laboratories.*)

Owing to their inherent frailty, it is necessary to protect the conductors of telephone cords against undue stresses, such as would otherwise occur from such simple accidents as dropping a receiver. Throughout its length the structure of the cord is intended to provide such protection, the outer braiding or the interior textile reinforcement, or both, being designed with that in view. The weak point of most telephone cords is at the terminals where the flexible conductor is joined to a rigid one. Here, the bending moments are most severe, and the textile reinforcements are less effective. Accordingly, some method of anchoring the cord at its ends in such manner as to minimize danger of breakage or injury at these points should be employed.

There are a number of these methods, but all of them depend upon the same principle, that of making a short tie between the textile body of the cord and the point of anchorage. The tie

is made so short that when it is drawn taut a considerable amount of slack will be left in the conductors leading from the body of the cord to the terminals.

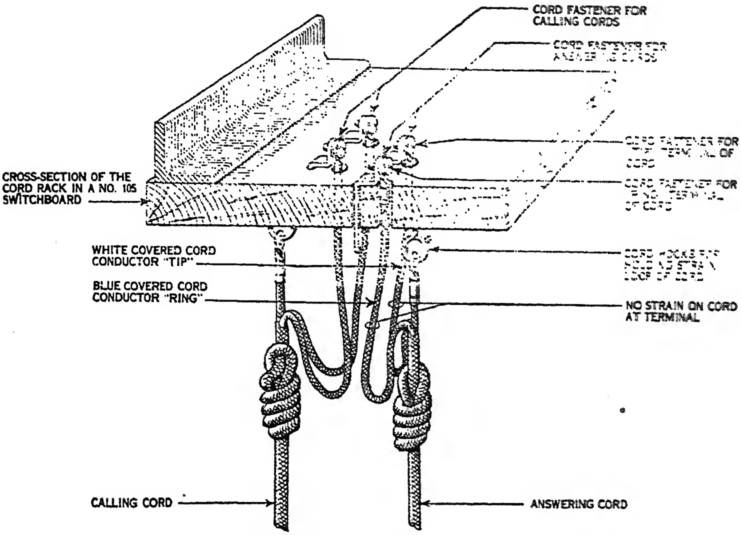


FIG. 246.—Rack end of cords on Stromberg-Carlson switchboard.

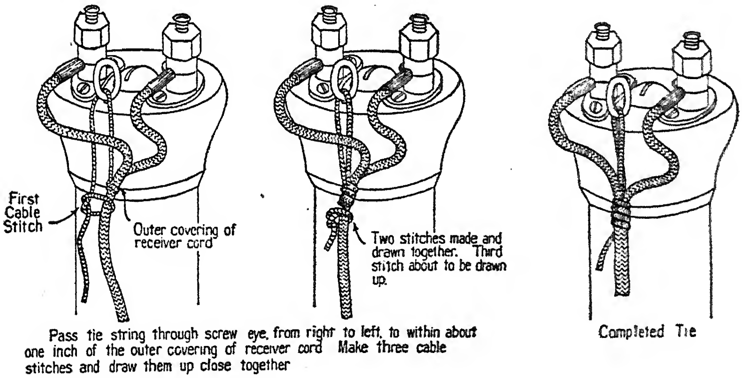


FIG. 247.—Tying cord on receiver with external binding posts.

One common method, almost universally used until recent years, has been to provide receiver, desk-stand and switchboard cords with "tie strings" or "strain loops" intended to absorb all pulls that would otherwise fall on the cord terminals. These usually form mere continuations of the outer braided covering

of the cord. When, during the braiding operation, the braiding machine reaches the point on the length of the cord where the outer covering is to stop, the conductors are taken out of the braid and the machine continued in operation until a string several inches long has been woven. Such a strain loop is shown on the switchboard cord of Fig. 242. Here, it is provided with a metallic eye to engage a supporting hook. The manner of its use and of attaching the cord terminals to the cord-rack terminals in a Stromberg-Carlson switchboard is made clear in Fig. 246.

Figure 247 shows the anchoring of a receiver cord with tie string to a former type of Bell receiver using external binding posts.

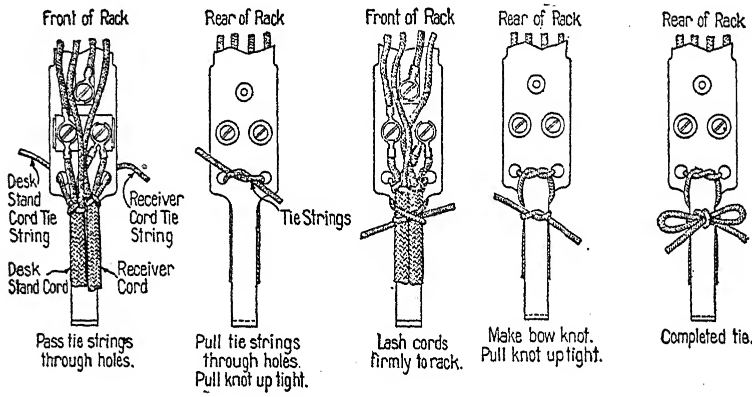


FIG. 248.—Tying cords to Western Electric desk stand.

Another instructive example of tie string fastening is shown in Fig. 248. In this the rack shown is the terminal rack extending down through the tubular standard of the Bell desk stand. The desk-stand cord and the receiver cord both enter a hole in the base of the desk stand, not shown, and extend up the rack to the terminals on its front. Here the tie strings of both cords are used to lash the bodies of the cords securely to the shank of the rack, so that any pulling on the cords themselves will be resisted by this lashing and not by the terminals.

Instead of tie strings flexible cords are sometimes provided with "stay hooks," small metal hooks securely anchored to the end of the outer braiding. A cord so provided and fastened to a Western Electric desk-stand rack is shown in Fig. 249.

Still another method that has come into use during recent years is to provide neither tie string nor stay hook but, instead,

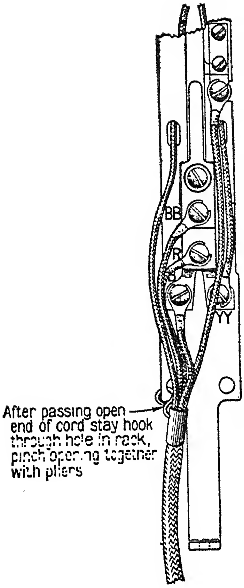


Fig. 249.—Desk-stand cord with stay-hook.

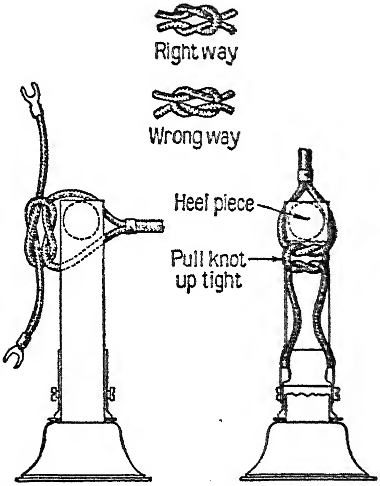


Fig. 250.—Tying receiver cord without strain loop.

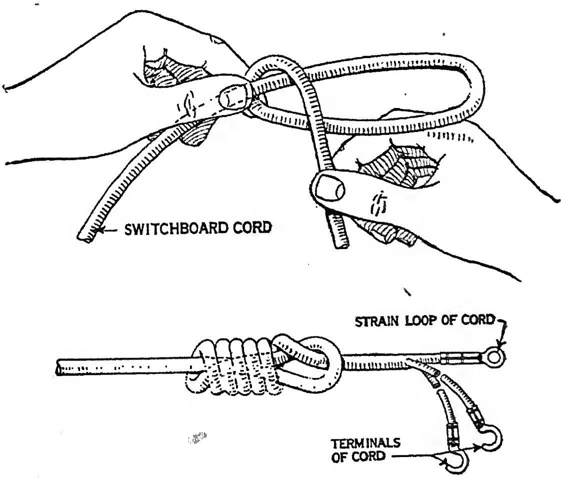


Fig. 251.—Taking up slack in switchboard cords.

to tie the conductors themselves in place, allowing sufficient slack beyond the tie to prevent any stress coming on the conductor terminals. Such a tie, applied to one of the present standard Bell receivers, No. 144, is made as in Fig. 250. The now generally adopted practice of placing all electrical connections inside the apparatus casing, in this case the receiver shell, marks a decided improvement in flexible-cord practice. The cord terminals are protected from accidental injury and from meddlesome or inadvertent tampering.

The knot in the body of each of the cords shown in Fig. 246 calls attention to another phase of switchboard cord practice. New switchboard cords are sometimes purposely furnished somewhat too long, and it is necessary to take up some slack in order to prevent the cord weight from striking the floor. The extra length is stored in this knot and later becomes available, as the cord is gradually shortened by cutting it back and reforming it at the plug end. The details of the making of this knot, as given by the Stromberg-Carlson Company, are shown in Fig. 251. Obviously, the knot may be used to store up any amount of slack required, according to the number of times the cord is wound around itself.

CHAPTER XVI

JOINTS AND CONTACTS

Before the advent of the telephone, telegraph workers had learned something of the difficulty of making good joints and contacts in their wiring and apparatus and something of the bad effects resulting from their failure to do so. Their state of mind on this subject was well expressed as early as 1878 by an English electrical paper,[1] as follows: "Electricians have long had sore reasons for regarding a 'bad contact' as an unmitigated nuisance, the instrument of the evil one, with no conceivable good in it, and no conceivable purpose except to annoy and tempt them into weakness and an expression of hearty but ignominious emotion."

The telephone man of today holds the same opinion of bad contacts. With him they are even worse than with the telegrapher, because of the very feebleness of the currents employed and the extreme sensitiveness of the receiver in detecting any abnormalities of current such as are often caused by faulty contacts. Fortunately, however, the technique that has been built up during the last fifty years has gone far toward their extermination. But only extreme vigilance can keep them down, and, even with that, they do occasionally arise to cause extra work for the telephone man and annoyance to the patron by interfering with his service.

The securing of a proper contact between conductors when contact is desired, and a complete separation when a break is desired, while, apparently, a small matter, is one upon which a vast amount of scientific research work has been expended and one which lies at the very foundation of good telephone service.

For the purposes of this chapter we may roughly classify those joints or contacts where two or more conductors are brought together to establish an electrical path between them as "permanent joints," "semipermanent joints" and as "make-and-break contacts."

[1] *Telegraph Journal and Electrical Review*, July 1, 1878.

Examples of permanent joints are found in splices connecting two wires with the intention of forming a permanently continuous conductor, or, in the soldered joints between wire ends and apparatus terminals made with the same intent.

Semipermanent joints are those which provide for the occasional intentional severing of the connection without injury to the conductors. Joints of this kind are made by means of binding posts, such as those at the terminals of dry batteries, or by means of various forms of "test connectors," all of which provide means for the occasional opening of the circuit, as for testing, for the replacement of apparatus, or for other purposes.

Make-and-break contacts are those which definitely function alternately to make and break the circuits with which they are associated. They may be operated manually or automatically. Examples of manually operated make-and-break contacts are found in the knife switch, the cooperating elements of plugs and jacks and in the various types of push buttons and keys. Familiar examples of automatically operated make-and-break contacts are those in the ordinary electric door bell and in the electromagnetic relay.

The classification of joints and contacts just referred to must be taken with discretion. A so-called "permanent joint" in a line wire may "go bad" and open the circuit. It has lost its permanence. It may even make and break the circuit with its swinging to and fro in the wind, and thus become a make-and-break contact. But these abnormalities do not alter the basis of classification. They are but diseases of a permanent joint that has gone wrong. Again, the soldered connection between a wire and an apparatus terminal is classified as permanent, although at times such joints are intentionally unsoldered and broken to allow the removal of defective apparatus. The principal intent of the soldered joint, however, is permanence, and it is in accordance with the principal intent that the foregoing classification has been made.

The making of permanent joints, such as between two lengths of bare wire, in open-wire construction, or between two lengths of insulated wire in cable work or interior wiring is a matter which more properly belongs in subsequent chapters, but for the sake of completeness here it will be briefly discussed.

Figure 252 shows the old so-called "Western Union joint," once largely used for galvanized-iron line wire. This joint, if

merely made as shown, will not develop the full tensile strength of the wire itself, nor is it entirely immune from troubles of defective contact as the corrosion of the iron progresses with time. If, however, the neck of the joint is carefully soldered, the joint will stand more tensile stress than the wire itself, and its permanent conductivity is greatly improved. The heat of soldering, however, is likely to injure the galvanized coating and induce corrosion. Such a joint, without soldering, may be made to

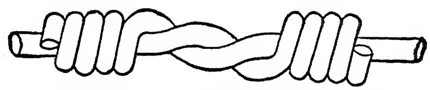


FIG. 252.—Old Western Union wire joint.

develop the full tensile strength of the wire by making its neck long enough to include five complete twists instead of about two, as shown.

The generally accepted way of joining bare hard-drawn copper line wire is by means of the McIntire sleeve joint shown in Fig. 253. The sleeve employed in making this joint consists of two parallel copper tubes joined together throughout their length, each having a smooth internal bore of such size as to fit the wires to be joined.

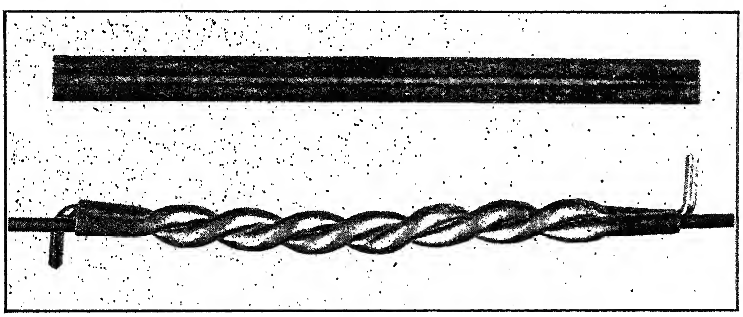


FIG. 253.—McIntire sleeve joint.

The two wires are run through the tubes in opposite directions and, then, by means of special clamps applied at each end, the sleeve is twisted through three complete turns. The ends of the wire are cut off about a half inch beyond the tube and bent over, as shown. Tensile strength is, of course, an important characteristic of hard-drawn copper line wire. The fact that this sleeve joint is effective without soldering is an important one in its favor, for the heat that would be required in soldering would

serve to anneal the wire and thus to destroy the strength gained by hard drawing.

Such sleeves are also used, under circumstances where soldering is impossible or objectionable, for joining smaller soft copper wires. They are also sometimes made of tinned steel for use in joining galvanized-iron and steel line wire.

An entirely different type of copper-wire joint (Fig. 254), also made without soldering, is used in splicing the conductors of lead-covered paper-insulated cables. The technique of making these joints and of their subsequent treatment in a cable splice is an important phase of cable work and will be described in detail in a subsequent chapter dealing with that subject. The joint *per se* may be briefly referred to here.

A paper sleeve about three inches long is first slipped over one of the wires to be joined and slid back out of the way. The two ends of the paper-insulated wire are then twisted together so as

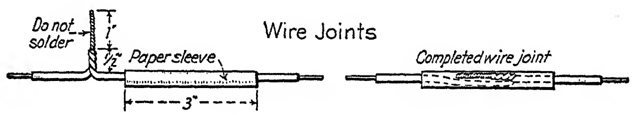


FIG. 254.—Cable-wire splice.

to include several turns of the insulation of both within the twist. This prevents the insulation from slipping back on the wires. Then the insulation is removed from both wires beyond the twist, care being taken not to nick the wire. The two bare wire ends are closely twisted together and cut off about an inch from the end of the insulation. The twisted portion is then bent down and the paper sleeve slipped back over it, as shown.

This wire joint is one of the examples in telephone work where it is permissible to join directly fine wires without the use of solder. The reason for this is that the wires, at the time of making the joint, are always clean and bright so that the long tight twist assures good initial contact and the joint is subsequently fully protected against corrosion, moisture and mechanical injury by "boiling it out" in paraffin and by covering it with an air-tight moisture-proof lead sleeve. Obviously, it is not, of itself, capable of withstanding much tensile stress or rough handling, but this is not necessary. The paper sleeves protect the individual joints from each other and the outer lead sleeve protects against mechanical derangement. Furthermore, the

methods employed in handling and finally supporting the spliced cable assures that the wires will be subjected to no great tensile stresses. Although not soldered these joints, when properly made and protected, give no trouble.

A fourth type of permanent wire joint is shown in Fig. 255. This is employed in working with various kinds of insulated wire, such as that used in house wiring and other interior work. The wrapping of the tails around the ends of the insulation prevents the latter from sliding back on the wire and also reduces the liability of the wires being kinked and broken at the ends of the more rigid splice. These joints should always be soldered, since they can be afforded no such protection against mechanical injury and corrosion as is given the unsoldered joints of a lead-

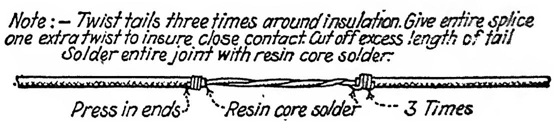


FIG. 255.—Splice in interior wire.

covered paper cable splice. After soldering, they are served with a wrapping of adhesive tape to afford some degree of insulation and mechanical protection.

In cases where an open flame to heat a soldering iron is objectionable and where no current is available for heating an electric soldering iron, joints in interior soft-copper wires should be made by double-copper sleeves, such as shown in Fig. 253; these, of course, being of proper gage to fit the wire. A joint so made should be taped to provide continuous insulation.

These four methods of directly joining lengths of wire in permanent fashion are, with slight variations, the principal ones used in telephony. They are not intended to be opened. If it becomes necessary to disconnect the wires, the entire splice is cut out.

We come now to the consideration of permanently joining wires to other things. Such joints are nearly always made by soldering the wire to a terminal of brass or German silver specially formed for that purpose and called a "soldering terminal." Joints of this kind are by far the most numerous of all the permanent connections employed in telephony, occurring literally by the million in large central-office installations.

The structure of soldering terminals and their arrangement with respect to the various pieces of apparatus of which they form a part are amply illustrated in many of the apparatus cuts in chapters to follow. In spring jacks, keys, relays and other apparatus employing contact springs, the terminal is usually a continuation of the spring, projecting beyond the body of the apparatus in a way convenient for attaching the wires. Such flat terminals may be notched or perforated for the reception of the wire, and they may occupy horizontal, vertical or oblique positions. The terminals of relay coils and other coils often take the form of stiff round wires projecting from the rear head of the spool, these being flattened and perforated at their ends to facilitate soldering. Some soldering terminals are made from straight stiff round rods, drilled longitudinally, or from small tubes, in which cases the wire is inserted longitudinally in the opening in the end of the terminal. This tubular type of terminal is becoming rare, however. Usually, it is no better and costs more than the flat type.

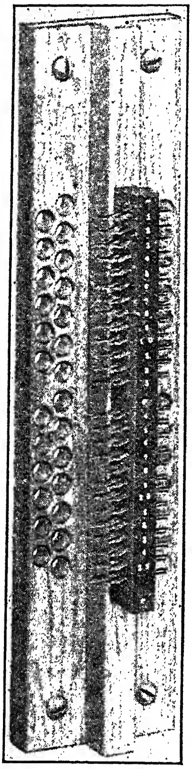


FIG. 256.—Terminal strips. (*Courtesy of Cook Electric Co.*)

In many cases the soldering terminal forms no part of other apparatus but is merely used as a convenience in joining wires together where large numbers of them have to be joined and arranged in orderly fashion. Large groups of such terminals are employed on central-office distributing frames. They usually consist of sheet-metal punchings arranged in convenient groups and subgroups and firmly mounted on insulating material. Such a group of terminals is called a "terminal block" or "terminal strip." A terminal strip of the Cook Electric Company is shown in Fig. 256.

As a rule soldering terminals are tinned as they come from the factory. This is done by dipping them into molten tin and shaking off the residue before it hardens. This greatly facilitates the subsequent soldering operation.

Another practice, which greatly facilitates soldering, is that of tinning the copper wire before its insulation is applied. This is

done by drawing the bare wire through a bath of molten tin and wiping off the surplus. The various types of wire largely used in central office wiring, such as *switchboard wire* and *jumper wire* should always be tinned. This is true whether the insulation on the wire consists merely of textile wrappings or of a combination of enamel and textile coverings, and it is true whether the wires are single, or in double- or triple-conductor combinations, or bunched into cables. When tinned, the wire always presents a good soldering surface, no matter where it is skinned for making a joint. The wires of lead-covered paper-insulated cables are not tinned.

The matter of soldering is one of the most important in the technique of telephone work. To be sure it is but a detail, but telephony is an art of almost infinite detail. To devote space to a subject so simple may seem trite, until it is stated that the soldering of the tinsmith or the electric light wireman will not do in telephone work.

The cardinal principle to be laid down in soldering small telephone wires to each other or to terminals is *to use no flux other than rosin.* All forms of acid flux and all the so-called "soldering compounds" and "soldering sticks" are to be avoided. There are no exceptions to this rule. This may seem a strange provision in the light of the well-recognized place of various soldering compounds in other fields of work. The reason for it is that these compounds usually, if not always, contain acid or some corrosive agent, which, if left on the joint even in minute quantities, is likely to cause trouble. It may do so by corroding the joint itself or by working back into the apparatus and there corroding and eventually eating off the finer wires and also ruining the insulation. Whatever chemical action occurs is likely to be aided by electrolysis in the presence of the currents carried by the wires.

The most convenient form of solder for telephone use is that wherein the soldering alloy is in the form of a hollow wire about one-eighth inch in external diameter. Within this slender tube of solder is included just the right amount of rosin to act as a flux. The result is that both solder and a safe flux are applied at the same time. A cross-section of one form of rosin-core solder is shown in Fig. 257. Here the tube of solder is pinched at short intervals to prevent the rosin from running out of or back into the tube when heated.

Given the proper materials, the prerequisites for successful soldering are a hot, properly tinned soldering iron and clean surfaces where the solder is to be applied. The making of soldered connections will be dealt with more minutely in the chapter on switchboard wiring, but, to make more complete the

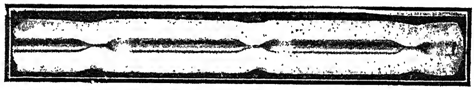


FIG. 257.—Rosin core solder.

general discussion at this point, Figs. 258 and 259 are presented with brief description. These show the general appearance of properly soldered joints on flat terminals of the drilled and notched types, respectively.

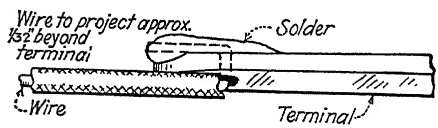


FIG. 258.—Soldered joint with flat perforated terminals.

After skinning the end, the bare wire is passed through the hole or laid in the notch of the terminal and wrapped about it sufficiently to hold during the soldering operation. The tip of the soldering iron is then applied to the connection at the same time flowing on a little solder, just enough to make a heat con-

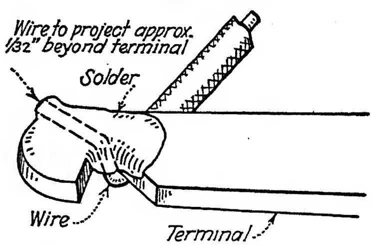


FIG. 259.—Soldered joint with flat notched terminals.

ducting contact. The iron is held at the joint long enough for the terminal and wire to become hot enough to flow the solder, the tip of the iron being rubbed slightly on the wire and terminal until the whole mass is thoroughly heated. The iron is then carefully drawn away from the terminal with a horizontal or downward motion but never by lifting it. In this way the excess

solder will follow the tip and may be applied on the next connection.

It is important to hold the iron on the joint long enough to assure the proper heating of the terminal and the wire, so that a complete union may be formed between their surfaces and the solder. It is also important not to overdo this, for if the heat is applied too long, the terminal may conduct enough of it back into the apparatus to cause injury there.. After the solder has become cool, the wire is clipped off close to it, leaving a finished joint as illustrated in Figs. 258 or 259.

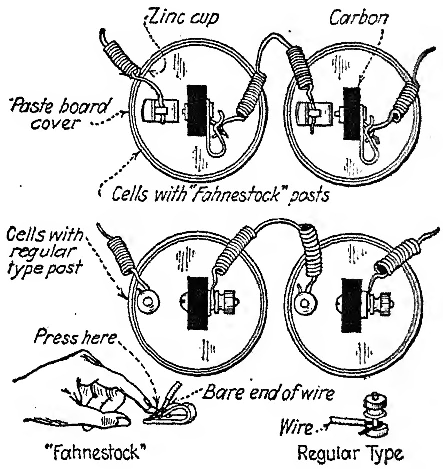


FIG. 260.—Two types of dry-battery terminals.

Semipermanent forms of joints or connections are usually made by some form of screw or lock-nut binding posts or by spring clamps. They are used in cases where the connection is occasionally required to be broken, as for testing purposes or for changing comparatively short-lived elements, such as dry batteries and receiver cords. They are also largely used in lieu of soldered connections where soldering would be inconvenient or objectionable, or where the skill and tools necessary to do good soldering are not likely to be available.

Familiar examples of two types of semipermanent connections as applied to dry battery cells are shown in Fig. 260. The Fahnestock type of clip is one of the most successful of many spring-clip connectors and has been used for many years for a

variety of purposes where convenience in disconnection as well as connection is desirable. Figure 261 gives useful hints about right and wrong ways of attaching wires to the ordinary screw and nut binding posts such as are shown in the lower part of Fig. 260.

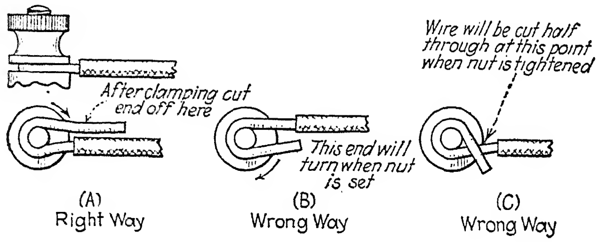


FIG. 261.—Ordinary binding post.

Fortunately, the once common type of binding post shown in Fig. 262 is falling into disuse. It is shown here to illustrate faults to be avoided. For the comparatively small soft-copper wires employed in telephony for interior wiring (No. 19 B. & S. gage and smaller) the screw of such a post is likely to be forced down so tightly as actually to cut off the wire, leaving a loose

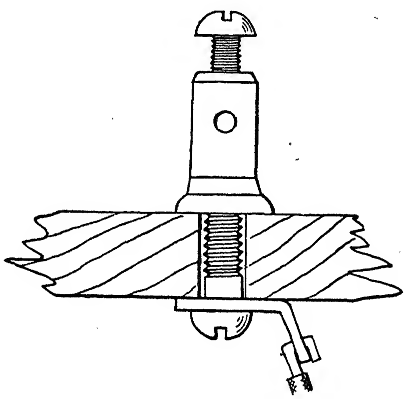


FIG. 262.—Faulty type of binding post.

contact with all visual appearance of a firm one. Again, the old practice of clamping the terminal of the permanent wire between the head of the mounting screw and the wooden board on which such posts were commonly mounted was objectionable. The shrinkage of the wood under varying conditions of moisture left loose contacts and also a loosely mounted post. Modern

practice is to connect the wire intended to be permanent, either by soldering or screw connecting, directly to the metal body of the post.

The lock-nut type of binding post has grown in favor. A group of them mounted on a single insulating base is shown in Fig. 263. This illustrates one of a great variety of terminal blocks made with various kinds and arrangements of posts to

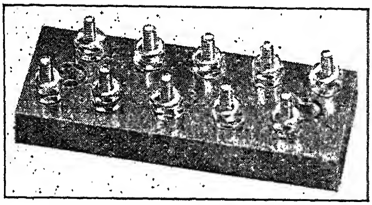


FIG. 263.—Connecting block.

facilitate connection, where a number of wires are brought together. In cases where some of the wires are to be permanently connected to such posts, the permanent connection is made by soldering the wire directly to the shank of the post which is elongated so as to project through and beyond the mounting block.

In other cases, the post has no permanent or back connections, merely standing as an insulated conductor for connecting what-

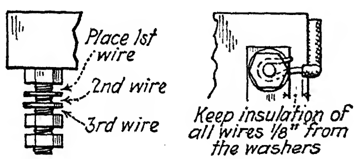


FIG. 264.—Lock-nut binding post for several wires.

ever wires may be brought to it. Such a post, of the lock-nut type, is shown in Fig. 264. A number of washers are supplied, depending on the number of wires to be attached. By placing the wires so that adjacent ones will always be separated by a washer, the danger of the wires cutting each other is avoided.

Still another type of connector, now largely used for telephone instrument terminals, is one where the wire, or the spade terminal of a cord, is directly clamped between the metal body of the connector and the under side of a screw head. These and the

method of their use with both wire ends and spade cord terminals are shown in Fig. 265.

A block of four such posts is shown in Fig. 266. Here, the bodies of the posts are molded in a block of phenol fiber, such as bakelite. They are connected in pairs horizontally, the two posts of a pair being in reality one piece of metal embedded within

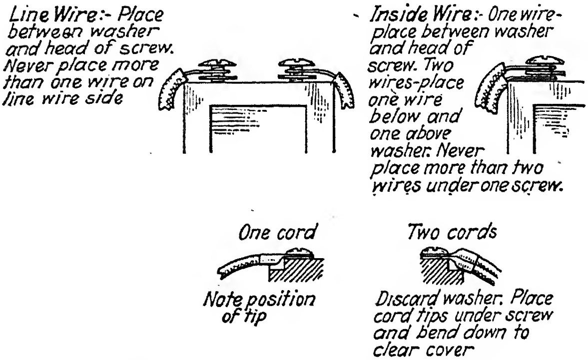


FIG. 265.—Connecting block for interior wires.

the insulating block. Such blocks as these are used, for instance, in terminating and connecting the inside and outside pair of wires of a subscriber's telephone installation. Only the inside pair in this cut is shown. To avoid accidental derangement, each pair of wires passes first through grooves on the under side of the block and is connected to the terminals on the far side.

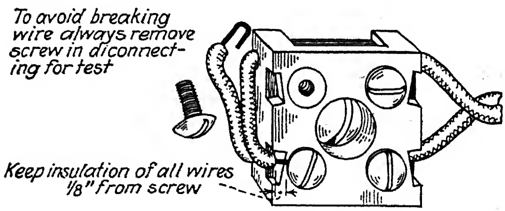


FIG. 266.—Connecting block for interior wires.

An entirely different type of contact problem is that of the so-called "make-and-break contact." In all of the foregoing permanent and semipermanent joints the main function is to keep the electrical path definitely and fixedly closed. In the make-and-break contacts, however, the purpose is just as much to

open the circuit as it is to *close* it. The cooperating contacts are alternately brought together to make the circuit and separated to break it, these acts being performed in the regular functioning of the apparatus with which the contacts are associated.

In telephony, chief interest in make-and-break contacts centers around the contact points employed in such apparatus as relays and keys, where circuits must be repeatedly made and broken, often with great rapidity and with small relative movements between the contacts and small amounts of energy available for causing the movement. The currents and voltages carried by the circuits which the contacts control are often so small that the slightest film due to corrosion, dirt or even a particle of dust lodged between the contacts may prevent the

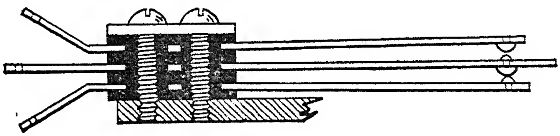


FIG. 267.—Simple spring pile-up.

flow of current when the contacts are supposedly brought together. On the other hand, the currents handled may be large enough, with enough electromagnetic inertia behind them, to cause destructive sparking when the contacts are broken. Such arcs as are formed may be minute, but, being concentrated on mere points of contact, the *density* of their energy distribution over the small areas involved may be comparatively great. Whatever sparking there is, of course, aids in the corrosion and ultimate deterioration of the metal.

Under heavy and long-continued sparking there is always some disintegration of the contacts, even with the best of contact metals. This is manifest by a pitting and roughening of the contact surfaces. Often a distinct "crater" will be formed by erosion in one contact, accompanied by an actual building up of the metal on the other. This will be referred to later after types of contact points have been discussed.

In Fig. 267 is shown one of the common arrangements of contact springs and their associated contact points. The springs are rigidly held, with respect to each other and their base, by screws and insulating separators. Their rear ends form soldering

terminals for connecting wires, while their free ends carry contact points, usually of some precious metal, either riveted or welded in place.

The switching arrangement here shown is one of the simplest, the long spring being adapted to make alternate contact with the two shorter springs as it is moved by some external force in one direction or the other. This would be referred to as a "one-make and one-break" combination. One group of two, three or more springs arranged in such manner is often called a "spring pile-up" and, as will be seen from the subsequent description of actual apparatus units, such as relays and keys, a single piece of apparatus may carry a number of such pile-ups variously arranged to make and break different contacts as their springs are moved into and out of engagement with each other.

Fig. 268.—Forms of contact rivets.

In the early days of telephony, after the need of precious metal contact was appreciated, the contact points were often soldered to their respective springs. Later the practice of riveting them in place was adopted and is still followed by many manufacturers. In about 1912 the Western Electric Company decided to attach its contacts by spot welding them to their springs and since about 1916 has done this, practically to the exclusion of other methods.

Contacts as ordinarily made are of two general forms, flat head and round head. One of each usually constitutes a cooperating pair. Their general forms, when adapted to riveting, are shown in Fig. 268. The dimensions indicated by letters in this figure are given in Table XV for a number of standard rivets as manufactured by Baker and Company of Newark, New Jersey, who deal extensively in electrical contacts of platinum, other precious metals and alloys.

The number of contacts to the ounce as given in this table are for pure platinum. For other metals the numbers per ounce would, of course, vary inversely as their respective specific gravities.

As will be shown, the principal metals now used for telephone make-and-break contacts are platinum, palladium and a certain

TABLE XV.—DIMENSIONS AND WEIGHT OF RIVET-TYPE PLATINUM CONTACTS

A Diameter, head, inches	B Thickness, head, inches	C Diameter, shank, inches	D Length, shank, inches	Approx. number to ounce
Round-head Rivets				
0.040	0.016	0.025	0.125	125
0.052	0.020	0.040	0.051	885
0.062	0.016	0.026	0.046	1,453
0.064	0.024	0.032	0.040	1,000
0.064	0.024	0.040	0.039	900
0.064	0.024	0.047	0.040	675
0.073	0.025	0.040	0.045	630
0.073	0.025	0.040	0.062	565
0.073	0.025	0.040	0.095	460
0.073	0.025	0.040	0.130	410
0.081	0.032	0.051	0.070	358
0.081	0.032	0.051	0.020	571
Flat-head Rivets				
0.050	0.015	0.030	0.040	1,500
0.062	0.025	0.032	0.025	980
0.063	0.060	0.040	0.050	323
0.070	0.031	0.040	0.015	637
0.075	0.028	0.041	0.030	548
0.080	0.012	0.032	0.044	925
0.080	0.050	0.040	0.050	251
0.080	0.060	0.050	0.060	210
0.081	0.032	0.051	0.043	339
0.081	0.040	0.040	0.050	391
0.082	0.031	0.041	0.034	420
0.090	0.005	0.028	0.028	1,880
0.091	0.040	0.055	0.028	263
0.093	0.031	0.062	0.035	260
0.093	0.035	0.040	0.032	326
0.094	0.031	0.062	0.095	174
0.095	0.010	0.038	0.031	915
0.100	0.006	0.040	0.039	900
0.100	0.006	0.047	0.038	750
0.100	0.031	0.051	0.044	285
0.100	0.032	0.041	0.046	280
0.100	0.042	0.050	0.055	200

platinum-gold-silver alloy. The other precious metals and precious-metal alloys used for electrical contacts to a greater or lesser extent are:

Pure gold.
Pure silver.
Alloys of platinum metals.
Alloys of platinum and iridium.
Alloys of platinum and gold.
Alloys of platinum and silver.
Alloys of gold and palladium.
Alloys of silver and palladium.
Alloys of silver and copper.

Before discussing the relative merits of the metals principally used, some of the methods in vogue for economizing the amount of precious metal will be referred to.

It is obvious that in the ordinary rivet composed entirely of precious metal, the shank of the rivet which extends through the contact spring is wasted, so far as useful contact purpose is concerned. Various methods of building up rivets of base metal with contact surfaces of precious metal have been proposed. In the one most widely adopted the entire shank and a portion of the head are of base metal. On this a solid precious metal contact disc or point is welded. Such a compound rivet is shown in Fig. 269. The saivng in precious metal is obvious, but the cost of manufacturing the rivet is greater than that of a one-metal rivet.

Fig. 269.—Composite contact point.

By welding the precious metal contact directly to the supporting spring, a process which, as stated, the Western Electric Company now uses exclusively, the required amount of precious metal is cut about in half. The process also eliminates most of the labor of making the rivet, reduces that of attaching it, and also minimizes the possibility of a defective joint between the contact itself and the supporting springs. A pair of contacts for welding directly to the springs consists of a flat disc and a point or button of the general forms indicated in Fig. 270. These are standardized in two sizes, as shown, which are termed, respectively, "light" and "heavy." The contact is cut from wire, formed and welded in one operation, without waste of metal and with great rapidity.

As illustrating the erosion of one contact and the building up of the other, and also the life that may be attained by contacts under certain conditions, Figs. 271 and 272 are given. These are photomicrographs, each magnified about fifteen times.

Figure 271 shows a pair of large or heavy type points of platinum taken from a Western Electric flat-type relay after

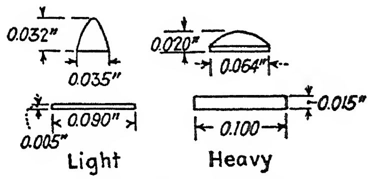


FIG. 270.—Welded type contacts.

it had operated in a laboratory test circuit about 44,000,000 times under conditions which gave heavy arcing. The erosion of the disc and build up of the point is to be noted.

Figure 272 is of a small or light type point and disc contact of "No. 1 contact metal" after approximately 27,000,000 operations in the same type of relay and test circuit as in the case of the

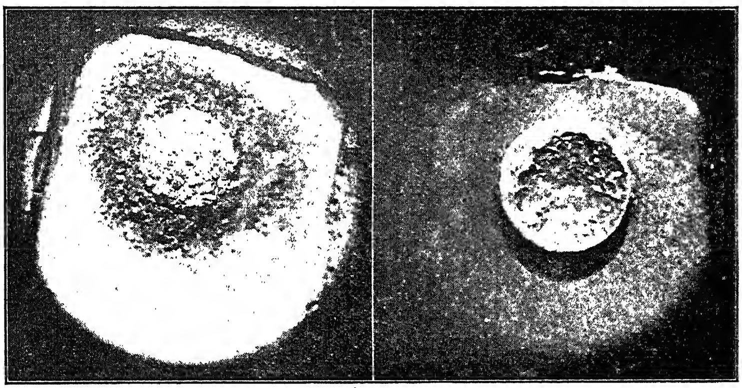


FIG. 271.—Heavy type contacts of platinum. (*Courtesy of Bell Telephone Laboratories.*)

heavy pair of contacts of the preceding figure. Here, the reverse condition is to be noted with respect to the direction of build up in that the point is eroded and the disc built up. The current in this test was apparently in the opposite direction. Also in this test the contacts were protected against heavy arcing by means of an electrical network bridged across the break.

Although a vast amount of research work has been done on it, the subject of make-and-break contacts for telephone work has not yet been exhausted, there still being much to be learned about it.

The use of platinum as the material for contact points is older than telephony. Its suitability for use in telegraph keys was recognized prior to 1860. With the coming of the newer art, platinum soon assumed its place as the ideal contact material, but its well-nigh universal adoption did not take place until after many grievous and costly experiences had been had with the use of baser metals. This deferred adoption was not surprising for, even in the early nineties, the cost of platinum was in the neighborhood of $20 a troy ounce, while copper, for instance—a far better conductor—could be had for as many cents a pound, and many of its common alloys for but little more.

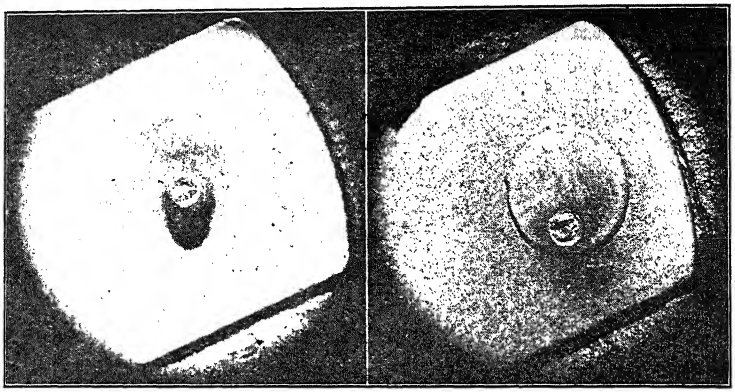


FIG. 272.—Light type contacts of No. 1 contact metal. (*Courtesy of Bell Telephone Laboratories.*)

The reasons for platinum's superiority over all other contact materials known at that time are several. Perhaps the principal one is that it does not oxidize in air at any temperature and is not acted on by any simple acid. It is practically non-corrosive under all conditions of use, which include not only ordinary atmospheric conditions but those within the electric arc. This non-corrosive characteristic minimizes the likelihood of the formation of a high-resistance film over the contact surfaces which would cause complete or partial failure of contact. A second important consideration is platinum's extremely high-fusing and volatilizing points. This minimizes the erosion of

the metal under the arc at times of sparking, and also the tendency of the points to become welded together under the action of currents of high density, such as often occur even with small currents when the actual area of contact approaches a mere point. Again, platinum is hard enough to resist excessive mechanical wear, is ductile and malleable enough to be easily worked, is a good conductor of heat, and, finally, has a fairly low specific resistance, which again contributes to comparative freedom from the welding together of the contacts.

The only thing to be said against platinum as the ideal contact metal is its scarcity, resulting in high cost and, at times, in uncertainty as to its being obtainable in sufficient quantity at any cost.

Its high cost became a factor of ever increasing moment as the number of contacts required in telephony multiplied with the growth of the industry, the increased demand for this use being the one of the causes of rapid rise in the price of the metal. Another cause contributing to this rise in price was the increasing popularity of platinum as a metal for jewelry. By 1925 the price had risen to the neighborhood of $125 a troy ounce and at times during the World War was considerably higher.

With characteristic foresight, anticipating the possibility of future prohibitive costs or even a failure of the supply, the American Telephone and Telegraph Company, as early as 1906, began an exhaustive series of experiments in search of a suitable substitute for platinum. Two major results came from the research. One was the invention of the so-called "No. 1 contact metal" of the Western Electric Company. This is an alloy of 5 to 7½ per cent platinum, 67½ to 70 per cent gold and 25 per cent silver. The other was the discovery that palladium, ordinarily a somewhat cheaper metal than platinum, was apparently a satisfactory substitute.

Before adopting either of these substitutes, they were subjected for a number of years to a series of tests, in both laboratory and field, to determine their relative merits with respect to platinum under all conditions of use. This preliminary work enabled the Bell system to meet a very serious situation which arose in 1915 and 1916 due to the shortage of platinum, brought about by Russia's participation in the war. It was then deemed safe to introduce the No. 1 metal and palladium extensively in manual systems. This successful use in actual practice fully justified the

conclusions reached on the basis of laboratory and field tests, and, as a result, the use of these substitutes has since been extended almost universally into machine-switching apparatus where the requirements are even more exacting. At the present time over 90 per cent of all the contact points which the Western Electric Company is supplying to the associated Bell companies are of the No. 1 metal, the others being mostly of palladium.

During the quest for suitable substitutes for platinum another scarcely less fruitful line of experimentation was going on in the Bell laboratories and shops. This was directed toward a saving in the amount of metal required in each contact. It resulted in the perfection of the system already referred to of electrically spot welding the precious metal contacts to the springs instead of riveting them, saving about one-half of the precious metal formerly required by one-metal rivets.

No better illustration of the practical value of wisely conducted scientific reasearch applied directly to industrial problems can be found than that just outlined in relation to contact points. It is the more striking because directed to such a small and apparently insignificant thing, a little piece of metal usually smaller than an ordinary pin head. Yet it is easily demonstrated that by the finding of cheaper metals and by the development of processes which conserve the amount of metal, a saving of over $30,000,000 was effected in the first cost of apparatus purchased by the associated Bell companies during the ten-year period ending 1926. The saving of investment in this item alone effects a saving in the annual charges of these companies of approximately $5,000,000.

At present the situation with regard to the choice of metals for contact points in telephone circuits seems to be about as follows:

Contacts of base metals such as copper, phosphor bronze and German silver are unreliable unless they are established with considerable pressure. and some sliding action between them. This is particularly true where voice currents are carried, because the slight potentials usually involved with such currents may be insufficient to break down the minute films that may form on the surfaces of the metal and thus prevent the establishment of the circuit. Sometimes such a contact will become "microphonic" and introduce puzzling noises in the telephone circuits.

On the other hand, we must avoid any sweeping condemnation of base-metal contacts for they play an important part in telephone work. Ordinary knife switches, the plugs and jacks

of manual switchboards and the wipers and cooperating contacts of automatic switches are all examples illustrating the fact that, under proper conditions of pressure and relative sliding action, base-metal contacts are thoroughly reliable.

Even in some relays, where a long powerful stroke of the armature is available, and where no great make-and-break speed is required, base metal contacts have been used with entire success. A notable example of this is in the 31-point relay of the North Electric Manufacturing Company described in a later chapter. In this multicontact relay each switch member consists of a long flexible phosphor-bronze blade which, when the armature is attracted, is forced between two cooperating German silver springs giving a knife-switch action.

As between platinum and palladium, the two are generally considered as fully equivalent, although some authorities hold palladium to be slightly inferior under some conditions. The only serious objection to palladium is its limited supply, it being obtained chiefly from working over the residues of old platiniferous workings. It probably could not be obtained at any price in sufficient quantities to supply the whole demand for contacts. Its specific gravity is about one-half that of platinum (Pd. 11.4, Pt. 21.5), a fact that is favorable to the cost of the palladium contact, since these metals are sold by weight.

As between platinum and the No. 1 contact metal (platinum-gold-silver alloy) ample experience seems to have proved that the alloy is as good for all cases except where comparatively high currents or large amounts of energy are encountered. There is some difference of opinion as to where and how the dividing line should be drawn between those cases where the No. 1 metal is as good as platinum and those comparatively few cases where platinum is the better. At present, the practice seems to be to limit the use of the alloy to current flows not exceeding one-half ampere and to amounts of energy at the break not exceeding 0.032 joule.[1] If either of these limiting conditions is exceeded, platinum is considered better. The current limitation is a simple one, but the energy limitation is more involved, requiring an analysis of the energy of the spark during the momentary period of its duration.

[1] The joule is the amount of work done by one watt in one second. Joules are represented by the product of volts, amperes and seconds or of watts and seconds.

Even platinum has decided limitations when much larger currents and large amounts of energy are encountered. Too much energy is involved per unit of surface, tending to shorten the life of the contact. Where circuits carrying more than one ampere are to be made and broken, the tendency is to use platinum-iridium contacts, which are very costly, or to double up the contacts or to take other measures to lessen the duty on a single pair of contacts.

While a completely satisfactory substitute for pure platinum has been found for use in perhaps 95 per cent of all telephone contacts requiring precious metals, there is no reason to believe that progress in this field has reached an end. Research work is being continued and there is reason to expect that other alloys will be found that will be even better than platinum, heretofore considered the ideal contact metal.

It must be remembered that the amount of sparking at the breaking of a circuit carrying a current is largely dependent on the characteristics of the circuit. At the time when the contacts are closed a certain amount of kinetic energy is stored in the circuit, principally in its electromagnet coils and other inductances. This energy, at least in part, must be dissipated at the contact gap when the circuit is opened, and it is the momentary flow of this "extra current" that is principally responsible for the arcing.

This fact that the destructive arcing is dependent on the circuit characteristics suggests the idea of possibly modifying the circuit to reduce arcing. This is in fact often done in one of three different ways: first, by putting a non-inductive resistance in shunt with the inductance; second, by connecting a non-inductive resistance in series with a condenser around the inductance; and, third, by connecting a condenser in series with a non-inductive resistance across the break.

The third of these methods of spark reduction is the one most commonly used. With it the transitory current following the break flows into the condenser instead of into the air gap. Having absorbed the energy that would otherwise be vented in the spark, the condenser prevents a continued flow after equilibrium has been established. The placing of a suitable non-inductive resistance in series with the condenser is important since without it the closing of the contacts would permit the discharge of the condenser through a circuit of practically no

resistance, which might result in such large current as to weld the contact points together.

The question of the size of the contact points is, within certain practical limits, one of relative economy. Ordinarily large contacts last longer than small ones, but they cost more. If a contact under a particular condition will be required to operate, say, 500,000 times throughout a reasonable expectancy of life, then, obviously, it is not worth while to use one capable of lasting through 5,000,000 operations. If, however, the number of operations throughout a reasonable life promises to be more than the contact will stand, then the question arises as to the relative initial cost of larger or better contacts, or of adding spark reducing condensers, as against that of the more frequent adjustments and renewals required by the inferior contact.

So far, those contacts designed to maintain an electrical path of constant conductivity between two conductors and also those which definitely function alternately to make such a path and to break it have been considered. Lying between these two distinct classes we must recognize the so-called "variable resistance contact" which functions in neither of these ways but rather to maintain a path of *variable* conductivity. Among these are some which persist in occurring in spite of all efforts to prevent them. The occasional bad contact, as in a faulty wire joint, or the make-and-break contact where a particle of dust or a high-resistance film has lodged between the points to prevent complete contact are examples. These must be looked upon as diseases.

On the other hand, the variable resistance contact, when kept in its proper place, serves one of the most useful purposes in telephony. The outstanding example of this is, of course, in the carbon transmitter where the purposely loose contacts existing between the carbon granules and the electrodes, and also between the granules themselves, are depended upon to bring about changes of resistance with the changes of pressure caused by vibrations of the diaphragm.

The variations in resistance through a defective wire joint, caused by the swinging of the wire in the wind, and those of a telephone transmitter, caused by the vibrations of the diaphragm acted upon by sound waves, are but manifestations of the loose contact phenomenon, working in one case *against* and in the other case *for* the desired purposes of telephony.

10432

INDEX

C

O

P

Q

V